Accessibility and Data Mining
Maximises Survey Research Potential

Margaret Kinsella

Blanchardstown Institute of Technology

2005

Mary Hendrick
Project Supervisor

Dr. Mícheál Ó hÉigeartaigh
External Examiner

Submitted in partial fulfilment of the requirements of the
Master of Science in Computing (ACCS), Institute of Technology, Sligo

Statement of Authenticity

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of the Masters of Science (Institute of Technology, Sligo) is entirely my own work and has not been submitted for an academic purpose other than in partial fulfilment of the requirements stated above.

Signed: Margaret Kinsella Date: August 2005

HETAC Registration No: S00018725

Acknowledgements

This project was supported by the Higher Education Staff Development Network and managed by the Institute of Technology, Sligo. Mr. Terry Young and his administrative team were very helpful throughout the MSc programme.

I would like to thank my project supervisor Ms. Mary Hendrick for all her help and support throughout this project. Mary has both a background and a keen interest in the research domain which extended clarity when validation was needed. I am sincerely grateful to her for her enthusiasm and dedication throughout the last year.

I would also like to thank, Dr. Mícheál Ó hÉigeartaigh, Waterford Institute of Technology, for agreeing to act as external examiner.

My thanks to colleagues in the School of Informatics & Engineering, particularly Geraldine Gray, Larry McNutt and Matt Smith, for their regular project research insights along the way.

Finally, extra special thanks go to my husband Martin and to my daughter Sinéad (her first year ran in parallel with this dissertation).

Acknowledgements (continued)

I would like to thank a number of people and organisations for their contributions:

Institute of Technology, Blanchardstown
Larry McNutt, Head of Informatics & Engineering
Diarmuid O' Callaghan, Registrar
Geraldine Gray, Colm McGuiness, Orla McMahon, Matt Smith (Lecturers)
Suzanne Walsh, Assistive Technology Centre

National BUA Centre & Dyscovery Centre, Wales
Dr. Amanda Kirby
Dawn Duffin
Suzanne McCarthy
Grainne Delahunt

Susan Bergin, Lecturer, NUI, Maynooth

Rob McCullagh & Habaca / SPSS Ireland

Sally Fincher, Head of Computers and Research Group, Kent University

Dr Chris Singleton, Dyslexia at Third Level Expert, University of Hull

Dr. Mark Magennis / Joshue O' Connor, Centre for Inclusive Technology, NCBI

Dr. Barry McMullin, Dublin City University, Web Accessibility Project

Myra O'Reagan, Data Mining, Trinity College, Dublin

Noel Redmond, Technical Support JAWS User, Creative Labs

Carlow IT

Information Society Commission

Irish Deaf Society

Irish Dyslexia Association

Jackson Assistive Technologies

National Council for the Blind

Royal National Institute for the Blind

Trinity College, Dublin

Visually Impaired Computer Society

Executive Summary

Project Title

Accessibility and Data Mining maximises Survey Research potential

Motivating Problem

This research project comprised an investigation of the process of designing and developing a generic survey research tool with an accessible user interface that will collect data for data mining.

Survey Research is a very specific method of collecting data. Accessibility (web accessibility and assistive technologies) is a core issue in ensuring a more inclusive and complete view of the user population being surveyed. A learning preferences assessment model was selected as a specific implementation/test case to illustrate proof of concept and to meet an existing college need. This specific implementation is currently in the form of a paper based assessment and when completed identifies personal preferences and difficulties for learning, to enable the student to be more successful in education. The individual is not limited by their preferences, but understanding them may help them be more productive and effective as a student. It is foreseen that the collective inputs integrate into a data mining tool, to produce results, which may inform college policy and decision making.

Aim of Research

The project's overall aim is to investigate the issues involved in designing a generic survey research tool in accordance with the following criteria:

- implements a learning preference assessment
- with an appropriate accessible user interface
- that will collect the data for data mining

Research Questions

1. How can the design of an accessible system address the problems of Survey Research data?
2. What are the optimal data mining considerations during the design of the survey framework (incorporating both structured and unstructured data)?

Research Hypothesis

The specific Survey Research implementation is a learning preferences assessment tool for the Institute of Technology, Blanchardstown. It is critical that the survey is accessible; accessible in terms of survey content, assistive technologies and user friendly interface. The survey content needs to be configurable and the survey content dynamically populated from a database while also storing student responses.

Data mining is very specific in the forms of data being presented as input to the data mining tool for analysis. The data mining algorithms usually dictate structured numeric fields; however, a Survey Research may contain structured and unstructured data. How can unstructured data be incorporated?

Therefore, this project will determine the key issues involved in designing a generic system which is accessible to most users to gather specific data which will later be analysed by data mining.

Research Methodologies

The research analysis is based on a combination of primary and secondary research; primary research in the form of identifying and interviewing key stakeholders and domain experts and secondary research by reviewing current theory and case studies.

A Learning Assessment Framework model to be created by the author, has been identified as the basis for Survey Research proof of concept. This Learning Assessment Framework endeavours to be accessible to all, including people with disabilities and is part of an initiative proposed by the BUA Centre, an organisation set up to address difficulties around learning.

The testing is performed by the developer and a focus group of diverse users testing accessibility issues.

The software development methodology chosen is an object oriented approach comprising of the following phases; analysis and design, implementation, testing and evaluation review. The data mining methodology CRISP-DM (Cross Industry Standard Process for Data Mining) is the standard used in deploying the data mining technology.

Thesis Organisation

Chapter One: Introduction

This chapter elaborates on the problem definition, providing additional background information on the key areas; Survey Research, learning preferences assessment, accessibility and data mining. The research project methodology is presented in detail illustrating the flow of work in addressing how each of the specific objectives of the project will be met.

Chapter Two: Research

This chapter presents the secondary research pertinent to each of the project key areas; Survey Research, learning preferences assessment, accessibility and data mining. The discussion on learning styles and difficulties is in the context of a learning preferences assessment. Current accessibility standards will be outlined followed by a case study pertaining to Irish web site accessibility. An investigation and analysis of data mining and the CRISP-DM methodology is then discussed. The issue of structured and unstructured data will also be introduced.

Chapter Three: Research Analysis

This chapter presents an analysis of the primary research specific to this project. A number of key stakeholders and domain experts were identified and interviewed and the core issues discussed are grouped together by key project areas; Survey Research, learning preferences assessment, accessibility and data mining. A case study is presented to evaluate the key failings of Irish web sites, to prioritise the essential accessibility elements. Web accessibility is one method of ensuring user inclusion but without adequate assistive technologies some of the target population may not be included. Key assistive technologies are selected to allow survey participation to users with disabilities.

The system requirements are presented with an overview of the proposed system. An analysis of the data is then presented and the main components of the system are described.

Chapter Four: Design

This chapter presents an overview of the system design and development. The system design is presented and the data warehouse like structures are described. The approach followed for the design of the system is outlined and the model is presented using UML. The implementation technologies are described and the main components of the system are described in detail.

Chapter Five: Evaluation & Testing

This chapter documents how testing was conducted in relation to the survey application framework and data mining. The test conditions in which the research objectives were evaluated are outlined.

Chapter Six: Conclusions & Recommendations

This chapter evaluates the project against the overall objectives as outlined in the Executive Summary. It also provides several conclusions and proposes future research and development opportunities.

The Appendices provide supplementary information with regards to learning theory, database content for the learning preferences assessment, accessibility legislation and guidelines, data mining and text mining.

The references and bibliography are listed.

Glossary of Terms and Abbreviations

AT	Assistive Technology
Bobby	Automated web Accessibility Tool
CFIT	Centre for Inclusive Technology
CRISP-DM	CRoss Industry Standard Process for Data Mining
CSS	Cascading Style Sheet
EDEAN	European Design for All e-Accessibility Network
ETL	Extraction, Transformation, Loading
HTML	HyperText Markup Language
Modality	Multiple Learning Modes
NALA	National Adult Literacy Agency
NCBI	National Council for the Blind in Ireland
NDA	National Disability Authority
RNIB	Royal National Institute for the Blind
JSP	Java Server Pages
SPLD	Specific Learning Disabilities
SQL	Structured Query Language
UML	Unified Modelling Language
VICS	Visually Impaired Computer Society
W3C	World Wide Web Consortium
WAI-DA	WAI-Design for All
WAI-TIES	WAI-Training, Implementation, Education and Support
WAI	Web Accessibility Initiative
WASP	Web Accessibility Support Project
WCAG	Web Content Accessibility Guidelines

Table of Contents

Table of Figures & Tables

1 Introduction

1.1 Chapter Summary

This chapter describes the rationale and context for the project and provides an introductory background to the project areas of generic Survey Research, learning preferences assessment, accessibility and data mining. The project objectives, project research approach and methodologies and the project high level plan are explained.

1.2 Purpose of the Study

Changing educational trends and access initiatives are including a more diverse student population in third level education; students with disabilities, disadvantaged students, mature students and students for whom English is not their first language. Colleges are faced with the challenge of establishing more effective student retention initiatives and required knowledge on how students learn and develop to their best potential; knowledge on the most effective teaching methodologies to reflect the needs of students; knowledge to support funding applications for new programs for marginalized groups.

Knowledge for the student comprises of an individual learning plan highlighting their strengths and suggesting techniques and strategies to address their limitations. A learning preferences assessment tool in the form of a generic Survey Research will identify and chart an individual learning plan for the student while collecting data for analysis through data mining. The results from data mining it is hoped will aid the need for new knowledge for the educational establishment.

The incorporation of accessibility features for people with disabilities during the user interface design is to establish inclusion and parity for people with and without disabilities when they use technology-based resources [1].

1.3 Background to the Study

There are several threads to this research project and this background outlines the various components, players and relationships commencing with the Survey Research and its inherent problems. A learning assessment tool is used as proof of concept for the research when investigating the optimal approach to designing Survey Research systems to maximize data mining. There is an existing college need for this tool as part of a collaborative venture and the partners are introduced. The existing learning preferences assessment process is reviewed while highlighting Specific Learning Difficulties. Why accessibility is important is then introduced and the role of Data Mining is expanded.

1.3.1 Survey research

Organisations conduct surveys to find out the characteristics, behaviours or opinions of a particular population. Survey research gains much-needed intelligence and creates more value, enabling better decision making and forcing political pressures. However, survey research projects can fail if they are implemented badly at any stage. Some of the common mistakes that lead to bad data are poorly asked questions, data that is dirty or entered haphazardly, and a process that takes so long that action cannot be taken. In order for Survey Research analysis to be successful it is imperative that the individual responses translate into meaningful information to get the most out of their data.

1.3.2 Research Approach

A Learning Assessment Framework model has been identified as the basis for Survey Research proof of concept. This Learning Assessment Framework endeavours to be accessible to all, including people with disabilities and is part of an initiative proposed by the BUA Centre, an organisation setup to address difficulties around learning.

1.3.3 Collaborative Partners

BUA (Building on Achievement) is an organisation based on campus at the Institute of Technology, Blanchardstown (ITB). BUA is a collaborative partnership between NTDI (National Training & Development Institute), The Dyscovery Centre, Wales and ITB. The BUA Partnership was set up to address difficulties around learning that lead to students failing to achieve their full potential and/or failing to complete a course of study.

1.3.4 Learning Preferences Assessment Process

The BUA centre currently offers a paper based Learning Assessment for all students to determine individual learning preferences / profiles and identify difficulties. There is a requirement to record the information in a meaningful manner in a self service way and to analyse the information that enables feedback for changes and improvements for students. The analysis will, it is hoped, aid college decision making, courses/modules, student access & retention initiatives, disability awareness and possible legislation & funding changes.

1.3.5 Role of Learning Preferences

Individuals have developed methods by which they learn best. For example, visual, auditory and kinesthetic are typical base classifications. Many paper based and computerised models exist but do they match accurately the individual with a detailed profile providing them with a plan to heighten their strengths and techniques to limit difficulties; pragmatic recommendations and assistive technologies. Dunn et al. [2] concluded that students with strong learning style preferences demonstrated greater academic gains when learning through their preferences than students who had moderate or mixed preferences.

1.3.6 Specific Learning Difficulties

The term 'specific learning difficulties' (SPLDs) is commonly used to refer to Dyslexia. Table 1.1 outlines other difficulties and their conventional labels as follows:

Specific Learning Difficulty Label	Prominent Difficulty
Dyslexia	Literacy & numeracy difficulties
Dyspraxia	Co-ordination difficulties
Attention Deficit Disorder (Hyperactivity) Disorder (AD(H)D)	Attention and concentration difficulties
Asperger's Syndrome	Social and communication difficulties.

Table 1.1: Specific Learning Disabilities (adapted from BUA materials, author, 2005)

1.3.7 Importance of Accessibility

It is estimated that one in ten people have a physical impairment that hinders them using a computer. This includes many older people (55+) who inevitably experience reductions in the power of their eyes, ears and fingers. As many as 25% of Irish adults are functionally illiterate and up to 8% have some form of dyslexia. For many students English is not their first language. Providing inclusive technology presents quite a challenge [3].

1.3.8 Role of Data Mining

Data mining knowledge discovery is very specific in the forms of data being presented to the mining tool for analysis. The algorithms usually dictate structured numeric fields; however, a survey usually contains structured and unstructured data. A trade off may exist between good user interface design, that is able to capture the relevant inputs and data, and enables appropriate analysis from data mining algorithms. This project will determine if it is possible to design a system which is accessible to users to gather data which will later be analysed by mining without significant compromises to any facilities.

This will involve determining the appropriate data mining tool which provide facilities to easily find patterns and trends in survey data and provide algorithms to include deposits of textual data alongside structured data. It also involves determining the impact of the data preparation on the design of the user interface for data collection and analyse the impact of the data mining design on the user interface and data items.

1.4 Research Objectives

The main objective of this research project is to investigate the design issues in developing a generic survey research tool with an accessible user interface that will collect data for data mining.

The underlying objectives, within the project are:

1. How can the design of an accessible system address the problems of Survey Research data?
2. What are the optimal data mining considerations during the design of the survey framework (incorporating both structured and unstructured data)?

1.5 Research Methodology

The following tasks outline the methodology requirements of the research as required to meet the research objectives in section 1.4:

- Understand and develop the Business and Data requirements of the BUA organisation
- Research data mining concepts, technologies and standards
- Identify appropriate technology tools:
  - A data mining tool for data preparation and analysis
  - A web authoring scripting tool(s) for data collection at the user interface
  - A Database Management System for a generic survey application incorporating configuration and data access
- Identify an appropriate sample population of data
- Determine the best way to quickly and easily find patterns and trends in survey data
- Apply the data mining preparation techniques to this sample data
  - Dataset descriptions, inclusion/exclusion rationales, data cleaning, derived attributes, merged and reformatted data
- Design a suitable technology framework
  - Incorporate data mining preparation considerations to the data and user interface
  - Incorporate accessibility features for people with disabilities
- Implement the framework components
  - Survey Administration incorporating Learning Preferences Assessment
  - Survey Data Collection
  - Data Mining Analysis
- Test the framework and web accessibility
- Conduct a Post Implementation Review

1.6 Project Plan

Start Date	Tasks	Status
Oct 2003	Information Gathering Process Commenced Interviews & Questionnaires and review current papers and theory underpinning the following areas: • Survey Research • Learning Preferences Assessment • Software Accessibility & Assistive Technologies • Data Mining, Text Mining & CRISP-DM	Complete
Jan 2004	Analysis and Design Evaluate and Select Technologies Develop Initial Database Prototype in MS-Access Initial Prototype Review	Complete
Aug 2004	Self educate in the following technologies and skills: • MySQL Database Management System • Clementine Data Mining software • JSP and Java Beans • Resin web server • DBWings	Complete
Dec 2004	Perform Data Mining Scenarios & document observations	Complete
Feb 2005	Configure web development environment	Complete
June 2005	Develop web-based Survey Application prototype	Complete
July 2005	Review and evaluate Research objectives	Complete
August 2005	Complete and present Thesis documentation	Complete

Table 1.2: High Level Project Plan (author, 2005)

1.7 Methodologies and Standards

The following is a list of the project methodologies and standards adhered to:

- This project follows an object oriented approach.
- Implementing CRISP-DM as the Data mining standard.
- Evaluating accessibility using the WCAG 1.0 guidelines.

2. Project Research

2.1 Chapter Summary

This chapter will provide an in depth look at a number of research areas. The initial research focuses on how technologies can support *learning preference profiling* in the domain of *survey research.* The chapter contains the key research aim that identifies appropriate *assistive technologies* and the issues involved in software *accessibility* which will inform the design phase to allow for maximum user inclusion. A further research area looks at how the design of an accessible web based survey application will address the *problems of Survey Research data.* This chapter includes investigations into the benefits that *data mining* technology can provide to the Survey Research application. It also incorporates the *optimal data mining considerations* to be applied during the design of the survey application.

2.2 Survey Research

Organisations rely on survey research to gather much needed business intelligence to find out the characteristics, behaviors or opinions of a particular population. They seek to answer specific questions about the surveyed topic related to 'why', 'who', 'where' and 'what' [4].

Survey research is a multiple step process with a clearly defined protocol at each step. In order to get reliable results from survey research, it is necessary to plan the project, collect data, access and manage the data easily and report the results. For the project to be a success, the results need to be shared with the decision-makers who can act upon them [4].

Survey Research is an involved process of which 'asking questions' to 'elicit responses' is just one phase. When survey research is performed correctly, it provides information that can be acted upon, based on good data. The information received from survey research enables improved decision making. It can help organisations meet their objectives, such as capturing more customers, retaining students or improving student learning [4].

However, a survey research project can fail if it is implemented badly at any stage. Some of the common mistakes that lead to bad data are poorly asked questions, data that is dirty or entered haphazardly, and a process that takes so long that action cannot be taken or the data becomes old [4].

In order to gather information, a questionnaire or survey is used to ask the sample population a set of questions. This can be done through an interview or self-administered surveys. Once the data is collected and analysed, the results can then be reported and deployed [4]. In surveys, samples of potential respondents are drawn to represent the target population of interest known as the sampling frame. The sampling frame in this scenario consists of cluster sampling where the population consists of first year students studying Computing and Social Care in ITB. It is envisaged that the sampling frame would be extended to all first year students. The choice of the sampling frame can cause practical difficulties and introduce bias, thus having major implications on the interpretation of the survey results. For example data mining survey data from prison populations only, may suggest trends at variance with the full time student body.

Survey research, according to SPSS, can be divided into seven steps, as outlined in Table 2.1. One of the aims of this research project is to demonstrate how technology can improve the survey research process at each of the steps [4].

Steps	Description
1. Planning and survey design	An organisation should have a clear statement of the purpose and goals of the survey. It should clearly define the survey population, sample size, method of data collection and determine how results will be used. Questions need to be defined. This research project will explore in detail the design of an online configurable survey application that will enable an organisation to define a survey and its related questions which can be deployed over the Internet for data collection.
2. Data collection:	The many methods of data collection include telephone, mailed

Steps	Description
	questionnaires, face-to-face interviews, email and web-based. This research will focus primarily on the web-based approach with specific needs around accessibility.
3. Data access	The objective of the data access step is to pass the data that is collected into analytical software for further processing. A Data mining solution has been chosen to provide this facility.
4. Data preparation	The goal of the data preparation stage is to get the data ready for analysis. The research will explore how data preparation can be achieved using database data extraction and/or data mining facilities.
5. Data analysis	Data analysis is about extracting useful information from the collected. Simple data analysis can be performed manually; however the research will explore in detail how data mining software can extract the most value from the data.
6. Reporting	Organisations need reports to show the results of the analysis. Reports can be prepared manually or in conjunction with analytical software e.g. Data mining solution.
7. Deployment	Organisations need to deploy reports to the right people to ensure that the best decisions are made.

Table 2.1: Survey Research Steps (adapted from SPSS, author, 2005)

Case Study: On-Line Surveys with Special Populations

Charleston College in South Carolina has researched the 'recent experimentation of on-line surveys' which is due to the 'quick ascendance of the World Wide Web', yet they clarify that 'few attempts have been made' to conduct 'online experimentation with special populations' [5]. They recently completed two online pilot studies.

Research Focus

The focus of the research in relation to the online survey application is as follows:

- Investigate the current availability of online survey solutions
- Work with other academics with similar interests
- Determine the appropriate technologies to use in the design and implementation of the survey application
- Determine the key factors in the data model
- Determine the key factors in providing an accessible user interface

2.2.1 Current online survey solutions

A representative sample of current online survey solutions was reviewed. They showed rich functionality in terms of survey configuration and web deployment. However accessibility had not been considered. This conclusion was as a result of reviewing their respective online demos [6]. In addition, the applications contained in-built analysis capabilities but lacked integration with data mining technologies and other 3rd party data sources.

2.2.2 Collaboration with other academics

The nature of research in third level education frequently involves Survey Research data through on-line data collection. Susan Bergin, a PhD Researcher at the Department of Computer Science, NUI Maynooth (Supervisor Professor Ronan Reilly) collaborated in identifying the issues involved in the design of a generic survey tool. As part of her research, Susan required survey data and the instruments she used currently were paper based. The target population was first year students studying software development on computer science courses at ITB and a variety of other third level colleges (academic year 2004 / 2005).

2.2.3 Investigate appropriate technologies

A key component of this research is to design and implement a set of technologies to support and backup the research questions. The technologies included in the scope of the investigation are outlined in table 2.1:

Technology	Area of Interest
Database Management System	Survey Application
Web Server	Survey Application
Web development language(s)	Survey Application
Accessibility Technologies	Survey Application
Data Mining	Trend & Prediction Analysis

Table 2.2: Technologies (author, 2005)

2.3 Learning Preferences

Each learner possesses major learning systems that they combine in various ways to produce a unique individualized learning style. The acquisition of knowledge, skills and techniques vary considerably for each individual. Many students have learned to adapt effectively across sensory modalities and across learning systems developing their best potential.

Kirby highlights that it seems reasonable to assume that the greatest amount of learning can result when teaching is provided to accommodate to a person's learning strengths [7]. The difficulty arises with students with specific learning disabilities who may not be able to adapt to a teaching style that focuses on their weaker modalities.

Individuals who have a Specific Learning Difficulties (SPLD) have a specific difficulty in the way they process information which impacts on their ability to achieve full potential. They learn differently which means that, often, the traditional teaching methods may not work for these individuals.

A review of recent Learning Style approaches is summarized in Appendix A.

2.3.1 Learning styles using observational criteria

In addition to using standardised instruments, learning styles may be identified to a certain extent through classroom observation. It should be noted that observation in itself may not be sufficient to fully identify learning styles, but the use of a framework for collecting observational data can yield considerable information and can complement the results from more formal assessment.

Observational assessment can be diagnostic, because it is flexible, adaptable and can be used in natural settings with interactive activities. Reid and Given have developed such a framework - the interactive observational style identification (IOSI). A summary of this is shown in Appendix B.

Kirby also firmly states that '*many learners with dyslexic difficulties tend to have difficulty shifting from one way of learning to another, and for them it is vitally important that their particular learning style is identified and addressed through instruction that matches how they learn'* [7]. Low self esteem from previous educational systems is the confidence opposite which is necessary for learning exploration. A thoughtful approach to learning-styles instruction can help all students find the most effective ways of learning for them. Learning-styles instruction, Kirby continues equips learners with the skills for life long learning.

Correctly implemented, profiling should form the core of each student's management of their own learning. [7]

2.4 Accessibility

2.4.1. Key factors for an accessible user interface

Current accessibility focuses primarily on web accessibility with limited consideration for the provision of appropriate technology solutions addressing assistive technologies requirements.

2.4.2 Significance of Accessibility

Internet technology holds tremendous promise to significantly improve access to information and services for many people with disabilities. Properly engineered web sites can interoperate with dedicated assistive technologies to flexibly address a wide range of disabilities. The key is in the design of web sites so that they facilitate - rather than obstruct access by users with disabilities.

Dr. ZhangXu (Zhangxu & Aldis, 2001) outlines the paramount importance of accessibility when asked about the impact of Internet accessibility to him.

"If anybody asks me what the Internet means to me, I will tell him without hesitation: To me (a quadriplegic) the Internet occupies the most important part in my life. It is my feet that can take me to any part of the world; it is my hands which help me to accomplish my work; it is my best friend--it gives my life meaning"

2.4.3 Accessibility Goals

A service delivered through information technology will be accessible to the widest possible audience if the following conditions are met:

- All users are able to perceive and understand the controls, instructions and outputs
- All users are able to reach and manipulate the controls, inputs and outputs
- The user interface is consistent across functions, devices and repeated use
- For users who still cannot use the service, an equivalent alternative service is available

2.4.4 Legislation

The accessibility of Information and Communication Technologies (ICT) and the services delivered through them are covered under a combination of Irish, European and International legislation as outlined in Appendix C. This assists in driving change in web development practices.

2.4.5 Web Content Accessibility Guidelines and Strategies

The W3C (World Wide Web Consortium) and WAI (Web Accessibility Initiative) is the recognised authority and central repository for the development of Web accessibility guidelines. They are the developers of the Web Content Accessibility Guidelines 1.0 (WCAG 1.0) which describe how to make web content and websites accessible to people with disabilities and is summarised in Appendix D.

The guidelines consist of 14 separate guidelines, each of which has an associated set of one or more *checkpoints*. There are a total of 65 checkpoints which are classified into three *priority* levels. WCAG *conformance* levels are then defined on the basis of these priorities (W3C, 1999, Section 5).

If these checkpoints are satisfied by a web site, they will ensure that it has a high likelihood of being accessible to the widest possible variety of users. This is good for the disability community; but it is also good for the general community of web users: it is well established that *universal* design frequently results in services that are more usable for *all*.

2.4.6 Web Accessibility Initiative – Design for All

Design for All aims to increase accessibility of the Web for people with disabilities in European Union Member States by supporting the technical and guidelines development work done at W3C WAI with educational and tools-related activities.

W3C/WAI is further building on the accomplishments of WAI-DA through WAI-TIES (Web Accessibility Initiative: Training, Implementation, Education, and Support). WAI-TIES activities include improvement and expansion of materials supporting accessible website development, technical training and best practices exchange, and supporting increased standards harmonization.

2.4.7 Web Accessibility Review

Having highlighted the significance of accessibility; introduced recent legislation regarding accessibility; outlined the W3C structures, standards and guidelines supporting accessibility; it

might be understandable to regard web sites as accessible, when it is quite the opposite. A useful case study was recently carried out in Dublin City University by Dr. Barry McMullin entitled the Web Accessibility Reporting Project (WARP) where Irish web sites were reviewed against WCAG 1.0 standards. The results of this research are useful in prioritizing the key problems with web site accessibility.

Case Study: Web Accessibility Reporting Project (WARP)

Ireland 2002 Baseline Study by Dr. Barry McMullin Dublin City University 2002

The WARP research documents a study of over 159 separate web sites, operated by Irish organisations, and spanning a wide range of activities, information, and services. These were assessed for a set of characteristics correlated with the WCAG guidelines.

> *The key results are that, of this sample, at least 94% failed to meet even the minimum WCAG-A standard; and 100% failed to meet the professional practice WCAG-AA standard. Furthermore, at least 90% of sites failed to meet minimal conformance with generic technical standards.*

This should be a "wake-up call" for government, for public agencies, for private companies, organisations and individuals. It indicates that considerable progress could be made quickly, and with comparatively little effort.

2.4.8 Common Web Accessibility Problems

Table 2.2 summarises all WCAG Priority 1 defects, across all sites, identified in this survey, ranked by the proportion of sites:

Diagnostic ID	Description	WCAG Checkpoint	Incidence (sites %)
g9	Provide alternative text for all images	1.1	90.56
g39	Give each frame a title	12.1	33.96
g38	Each FRAME must reference an HTML file	6.2	33.33
g240	Provide alternative text for all image map hot-spots (AREAs)	1.1	26.41
g10	Provide alternative text for all image-type	1.1	18.23

	buttons in forms		
g21	Provide alternative text for each APPLET	1.1	10.69
g20	Provide alternative content for each OBJECT	1.1	0.62

Table 2.3: Most Common Accessibility Problems with Irish Web Sites (adapted from WARP, 2005)

Analysis of Common Problems

It is notable that 5 out of the 7 defect types in table 2.3 all refer to a single WCAG Checkpoint, 1.1: "Provide a *text equivalent* for every non-text element" [9]. Given the dominant impact of this category of defect on the overall results of the survey, it is worth discussing this in a little more depth.

The motivation for its use is elaborated as follows:

> *The power of text equivalents lies in their capacity to be rendered in ways that are accessible to people from various disability groups using a variety of technologies. Text can be readily output to speech synthesizers and braille displays, and can be presented visually (in a variety of sizes) on computer displays and paper. Synthesized speech is critical for individuals who are blind and for many people with the reading difficulties that often accompany cognitive disabilities, learning disabilities, and deafness. Braille is essential for individuals who are both deaf and blind, as well as many individuals whose only sensory disability is blindness. Text displayed visually benefits users who are deaf as well as the majority of Web users.*

Text equivalents thus provide a generic mechanism for addressing a wide variety of both web functions and user disabilities. The general notion of a text equivalent is usually pertinent in the context of providing alternatives to embedded images; but the scope of this checkpoint is significantly broader than that. Text equivalents can be an effective accessibility technique for all of the following web design situations; images (including those used as list bullets, graphical buttons etc.); image map regions; graphical representations of text (including symbols); animations (e.g., animated GIFs); scripts and applets; frames; audio and video content etc.

2.4.9 Cascading Style Sheets

Semantic HTML significantly reduces 'aural garbage' for the user interfacing with the GUI with a speech system [10]. An aural browser ignores the visual formatting properties defined in the Cascading Style Sheet (CSS), so the user need not be frustrated listening to properties. In real terms for a visually impaired user they would never have to wonder if a word was **bold** because it was more important or just because it looked better that way. Elements that were displayed in **bold** for design purposes would have that property assigned using CSS, and the aural browser would never mention it. Elements that needed additional impact would be marked up using the semantically meaningful and tags, which are displayed by default as bold and italic in visual browsers. Guideline 3 addresses the idea of avoiding presentational markup in favour of semantic markup [9].

This centralised design information ensures that the .css files house all the style code and is then linked by the appropriate pages with the HTML <link> tag. With this approach, everything to do with the *look* of the site can be found in one place, and is not interspersed with the *content* of the site. This has significant impact for the user experience and the bandwidth costs. This code *decoupling* also assists accessibility in that if a style needs to be modified, there is only one definition for these headings, instead of each HTML file.

2.4.10 User Interface Design Issues

A number of design considerations were identified and are illustrated in table 2.3.

Design Issue	Key Points
User Interfaces	Consistent layout and navigation. Allow user to choose own style sheets
User Control	Minimal inputs
User Workload	Optimal Interaction
User Actions	Consistent
Shortcuts	Compatible with main user agents
Error & Help	Consistent, clear, varying levels of help Error numbering for reporting
Browser	Compatible with modern browsers

Design Issue	Key Points
Response Times	Factor in download speed on remote machines
Graphics	Alternative text
Colour	• Selection of appropriate colour combinations • Style Switcher enabling user profile selection • Default sharp contrasting for Visually Impaired • Default pastel for people with cognitive • How best to cater for contradicting default colour schemes
Speech	Many new CSS Version 2 compatible speech properties exist for use by aural (speaking) browsers for the visually impaired. e.g. *Speak.* This property controls if and how an element's content should read aloud. *Speak-header* Controls how table headers are read Not yet supported by any currently available browser.
Separating Content from Presentation	Cascading Style Sheets

Table 2.4: User Interface Design Issues (author, 2005)

These considerations are applied in the practical design of the user interface. See Appendix L for further recommendations.

2.5.11 Assistive Technologies

For seamless functionality, usability based on a thorough understanding of users' needs should be a primary concern. Web accessibility alone will not ensure inclusive survey participation. Assistive Technologies are necessary to widen Survey Research inclusion.

Higham [11] thought it very necessary to persuade developers to think harder about usability i.e. to identify and understand users' expectations, characteristics, limitations and needs, as well as the task functionality and environment. For this research project this involved stepping back to consider a more diverse target population. The assistive technologies and their users in table 2.4 were researched as appropriate software tools / hardware devices necessary to enable a wider student population.

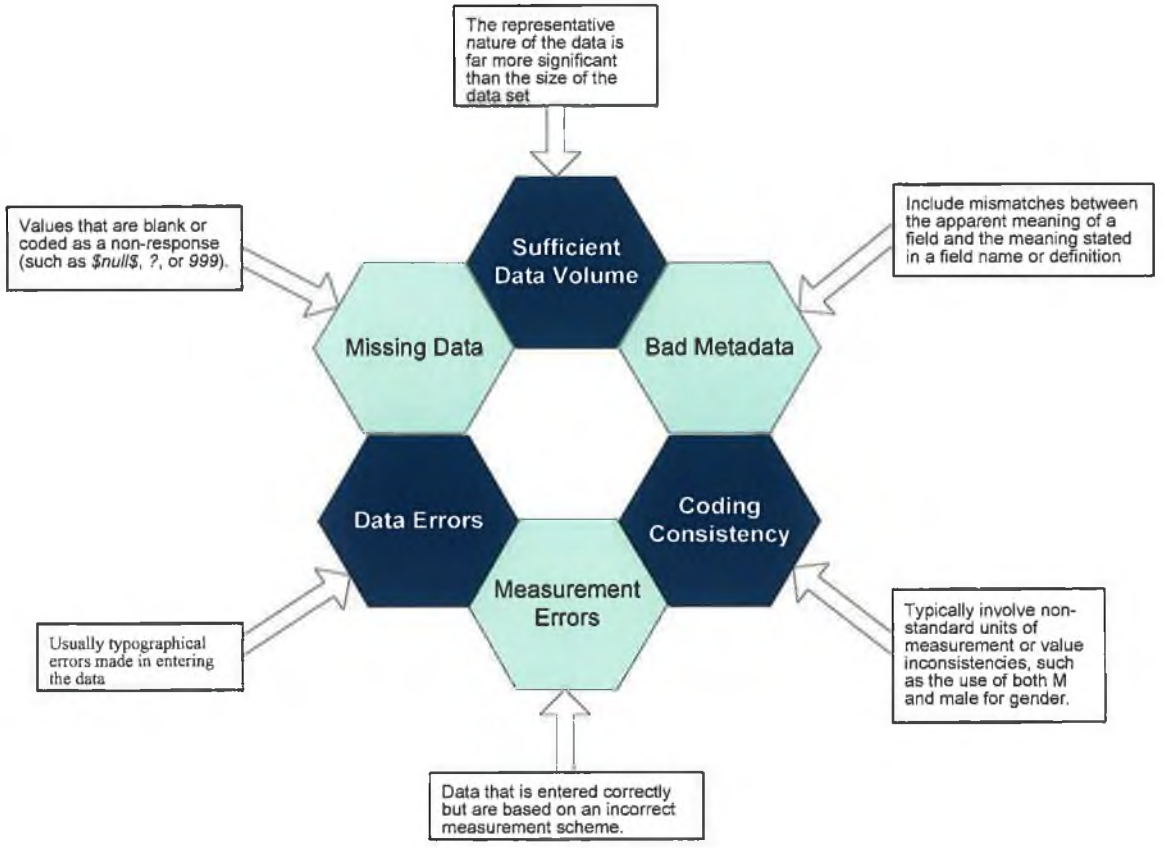
Disability	Accessibility	User Agents
Visually Impaired	Text Magnification	LUNAR ZOOMTEXT
Blind	Screen Reader	JAWS WinVision
Specific Learning Disabilities Dyslexia Reading Difficulties	Text to Speech	Webspeak
Physical Disabilities	Devices	Trackballs Joysticks Adaptive mouse inputs Adaptive keyboards Switch devices & software Access Devices Touch Screen
Deaf and Hard of Hearing	Proof English	

Table 2.5: Assistive Technologies and their uses (author, 2005)

2.6 Problems with Survey Research data

Data is rarely perfect. In fact, data may contain coding errors, missing values, or other types of inconsistencies that makes analysis difficult at times. Irrespective of the type of analysis tools used, it is very important that the potential pitfalls are avoided. This rule applies to considerations performed during the design of the survey application and equally during the data preparation process for data mining.

Figure 2.1 highlights six major areas where data problems occur. These are explained in further detail below.



(Diagram created by the author, 2005)

Figure 2.1: Problems that arise when collecting Survey Research data

2.6.1 Missing Values or Data

Missing values are values in the data set that are unknown, uncollected, or incorrectly entered. Usually such values are invalid for their fields.

For example, a field such as *Gender* should contain values such as *M* and *F*. If it is discovered that the field has values of *Y* or *Z* in the field, it is safe to assume that such values are invalid. Likewise, a negative value for the field *Age* is meaningless. Frequently, such obviously wrong values are purposely entered or left blank during a questionnaire or survey to indicate a non-response. At times it is important to examine these blanks more closely to determine whether a non-response, such as the refusal to give one's age, is a factor in predicting a specific outcome.

It is important therefore to restrict occurrences of missing values where possible in conjunction with the web development tools.

2.6.2 Data Errors

Errors found during the exploration process can be corrected. For the most part, though, a proper data entry facility should be enforced in a survey application before a student or respondent submits an answer to a back-end database. It should be the intention of any survey application to ensure that a large majority of the potential data errors be eliminated during its design and implementation.

2.6.3 Measurement errors

A poorly worded question on the survey or questionnaire can greatly affect the quality of the data. In such events, there may not be time to collect answers to a new replacement question. One of the aims of the survey application is to design a system that allows online surveys to be user defined and flexible to change. This allows for question updating in the event of discovering misleading or ambiguous questions.

In circumstances where problem questions are discovered after the event then the best solution may be to go back to the selection process with the data mining software and filter these items from further analysis.

2.6.4 Sufficient Data Volume

This is a difficult question to answer. For data analysis, it is not necessarily the size of a data set that is important. The representativeness of the data set is far more significant, together with its coverage of possible outcomes and combinations of variables.

For most modeling and analysis techniques, there are trade-offs associated with data size. Large data sets can produce more accurate models, but they can also lengthen the processing time. Consideration should be given to the possibility whether to use a subset of data. Typically, the more attributes that are considered then the more records that will be needed to give representative coverage.

2.6.5 Bad Metadata

Bad metadata occurs when there is a mismatch between the apparent meaning of a field and the meaning stated in a field name or definition. This is particularly so when creating survey questions where the person asking the question is assuming the respondent understands the question fully.

In many cases, the only way of identifying these kinds of problems is to manually examine suspect fields and track down the correct meaning.

2.6.6 Coding Inconsistencies

Coding inconsistency is when different data represents the same thing. Typically it involves non-standard units of measurement or value inconsistencies, such as the use of both M and male to describe a field called gender. Another example might be where an Insurance company has a field called "car type" in their system. A sample of the data entered include 'Merc', 'Mercedes', 'M-Benz', 'Mrcds'.

2.7 Data Mining benefits

Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills and academic domain knowledge to uncover hidden trends and patterns. These trends and patterns form the basis of predictive models that allow analysts to produce new observations (e.g. about students) from existing observations. The benefits of data mining are its ability to gain deeper understanding of patterns previously unseen using current available reporting capabilities [12].

The data mining definition from the Gartner Group seems to be the most comprehensive, as they define data mining as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques" [13]. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining finds patterns that are present in the data [12].

The trends and predictions that are generated from a survey can be useful from two perspectives. Firstly, the individual who completed the survey might be interested in knowing the information discovery from their own viewpoint. Also, the combination of survey responses will have particular interest to academics, managers and a wider audience. In the example of the learning preferences framework prototype, the use of information gathered and subsequent analysis could lead to knowledge discovery in areas such as those outlined below:

- The exploration of effective course design
- Study and analysis of student demographics (educational institutions could use to determine target markets for offerings)
- Evaluating and informing best teaching practices
- Analysis of course relevance to student needs
- Exploration of student usage of course materials within a course. This could be measured by the frequency of visits to material, numbers of FAQs on materials etc.
- It is possible that data mining would be able to reveal trends and patterns as learners participate in dialogue and interact with the learning environment.

- Potentially instructors and learners could be alerted to items that require attention in a course.
- Data mining could help learners achieve their learning goals, and help educators achieve their instructional goals.

Spreadsheets and databases provide simple summaries and basic row-and-column calculations. To best interpret and understand survey responses, in-depth analysis is required that is invariably not available in spreadsheets and databases. With data mining, survey responses can be translated into meaningful information and thus gain more insight into the responses.

Unusual responses can affect the results of survey analysis which can influence the decisions that are fed back to individuals and further group analysis. It is important to know whether an unusual response is the result of a data entry error, and should be corrected, or whether it reflects a true relationship that exists in the data.

One of the benefits that data mining has is that it helps to easily spot data entry errors, respondent errors or unusual responses that may want to be omitted from the analysis, or examined more closely. Data mining tools have several graphical facilities e.g. scatterplot that provides an overview of the data, helping to draw preliminary conclusions about possible relationships. It also helps identify "outliers" which can require closer examination.

For many reasons, students or respondents do not always answer every question. Missing responses occur when a question is not applicable, a respondent refuses to answer, or the respondent simply doesn't know the answer. Gaps in the data, influence the analysis and in turn, the results.

2.7.1 Data Mining Strategies

The types of knowledge or strategies for mining data include the following:

Characterization	Discrimination	Prediction
Classification	Clustering	Association analysis
Sequence analysis		

Each of the above strategies is now described in more detail:

Characterisation [14] is the discovery of rules that characterise (or summarise) some data.

For example one could characterise (or describe) graduate students in the science faculty. A rule (fictive) could look like:

IF Age in [20..35] AND Gender=M and Status=Foreign THEN Degree=Grad AND Spec=Science (85%)

85% here means that 85% of grad students in Science are between 20 and 35, males and foreign students. In other words, the rule holds in 85% of the known cases. These types of rules could be useful for education administrators. One could also characterise for instance students who succeeded in an exam. The characterisation could be based on attributes such as age, gender etc. but also attributes such as time of access, web pages accessed, participation in labs, etc.

Discrimination [14] is the discovery of rules that distinguish between two or more sets of data.

For example comparing graduate students in science and graduate students in Arts, one could find this (hypothetical) rule:

IF Degree=Grad AND Spec=Science THEN Status=Foreign(83%) while IF Degree=Grad AND Spec=Art THEN Status=Irish (89%)

One can also use these discriminant rules to compare for example access to a certain module by two different groups of learners for example "local" and "remote":

Access to Module 1 in the evening (Local 15%, Remote 85%), Access to Module 1 during the day (Local 70%, Remote 30%).

Here are some examples:

- Characterise one particular learner.
- Characterise a group of learners
- Characterise a conference or a discussion

In other words, what differentiates a particular learner from the others, a particular group of learners from the others?

Prediction [14] is the discovery of rules used to forecast missing data or future data.

Data prediction predicts some unknown or missing attribute values based on other information. Predicting the outcome of future events from a set of inputs.

At the University of Alberta [14], they have experimented with a program that predicts the next web pages that the user (or learner) would want to see and offer a list of potential links (with some probabilities associated to them). These become like a ranked list of suggested shortcut links to potentially useful resources. The prediction is based on previous visits of similar learners on the site. If the prediction is right, the user saves time jumping to the resources rather than drilling through the site to eventually finding the same resources.

Classification [14] is the discovery of rules that group data in given categories.

Classification is a kind of prediction. However, classification is typically for categorical data while prediction is for numerical data. What's the difference? Categorical data is for instance "red", "book" "rain", "[10..25]", etc., some arbitrary value belonging to a set, while numerical data are values such as continuous numbers: 20, 2.5, 55.99999, etc. Other points to note include:

- Classification is also known as supervised learning

- Provides a mapping from attributes to specific groups
- The groups are typically determined beforehand
- The variable that is used to determine the group is called the dependent variable
- The objective of mining the data is to predict the value of the dependent variable

An example to be considered for learning preferences might be if we want to classify learners based on their previous activities, their performance in some activities, the time they access, the time they spend on some activities, etc. The classes could be for instance "learner needs attention and monitoring" and "learner is autonomous and on a good track".

Any information about the learner, his/her performance in some activities, his/her access history, etc. could also helpful. For example, this is where integration with a 3rd party data source e.g. Examination results, attendance database, etc, could be very useful.

Clustering [14] is the discovery of groups in data, where the grouping is not known.

Clustering is the process of partitioning a set of data (or objects) in a set of meaningful sub-classes called clusters. It helps to understand the natural grouping or structure in a data set. The key concept in clustering is "similarity" or "dissimilarity" since a cluster is a collection of data objects that are similar to one another and thus can be treated collectively as a group. Clustering analysis, also called "unsupervised classification" because unlike classification there are no predefined classes. Other points to note include:

- Clustering is also called unsupervised learning because unlike classification there are no predefined classes.
- The data mining tools groups rows of data that share similar trends and patterns
- Clustering divides the data into specific groups
- There is no dependent variable (field), which means that splitting the data into groups is not necessary
- There is no upfront prediction of the outcome
- The objective is discover patterns in the data

They are using clustering analysis the University of Alberta [14] to discover groups of similar learning behaviours. In this case, they call learning behaviour as a sequence of activities in a

web-based learning site. They are experimenting by grouping web sessions in groups of similar behaviours, they not only group the learners who have similar activity patterns but also discover the existing learning behaviours. They are still investigating the important attributes that should be taken into account when grouping sessions. In other words, what constitutes a good similarity measure in the context of grouping on-line learning behaviours? Currently, they have only taken into account the sequence of events (i.e. the clicks visitation stream) per session. So a learner can end up in different groups if the learning behaviour changes.

Association analysis [14] is the discovery of rules that express associations between data.

These could be, for example, the items that are bought together in a transaction. The typical use of association rules is in market basket analysis. Given the transactions in a store for example, a store manager would want to know the strong associations between items bought. If an association rule is discovered between 2 items, item A and item B, it would be futile to have a sale on both items at the same time. A sale on item A would suffice because if a customer buys A, he or she will highly likely buy B as well.

For the learning preferences framework, an example might be where a student achieves two activities, a third learning activity might be recommended if there is an association rule discovered between the three activities based on historical data (data cumulated in previous or current surveys).

Sequence analysis [14] is the discovery of rules that express associations between sequences of events.

The idea is basically the same as association analysis but the order of the items matters, hence the sequence. So the items are not just in sets but in sequences. This is more important in the context of e-learning since the fact that a learner did an activity before another is relevant. The information "learner X visited page A after page B" conveys more knowledge than "learner X visited page A and B".

2.7.2 Data Mining Models and Algorithms

A model is an all-encompassing method that includes the steps, modules, and resources of the data mining process. A data mining model may include the entire process developed for a particular purpose, be it to cluster or predict. However, a model in data mining is different from an algorithm. An algorithm is a technical term to describe a particular mathematically driven data mining function. There are a number of algorithms for performing the different data mining strategies outlined above. Table 2.6 contains a list of some of the commonly used algorithms found in data mining applications to-day. In addition, the most prevalent strategy that pertains to each algorithm is nominated. Note that there are other types as more and more algorithms are being developed both commercially and by academics as different requirements arise.

Strategy	Algorithm
Prediction	Neural Net Node
Classification	C5.0 Node
Clustering	Kohonen Node
Prediction	Linear Regression Node
Association Analysis	Generalized Rule Induction (GRI) Node
Association Analysis	Apriori Node
Clustering	K-Means Node
Classification	Logistic Regression Node
Clustering	TwoStep Cluster Node
Classification, Prediction	C&R Tree Node
Sequence Analysis, Association Analysis	Sequence Node

Table 2.6: Commonly used data mining algorithms (author, 2005)

2.8 Optimal data mining considerations

Organisations typically embark on data mining assignments on well established datasets. They include databases and data sources that invariably were not designed with significant data mining considerations. However that is not to say that these data sources are not suitable for data mining. In many cases the database design contains a richness that is well suited to data mining while in other scenarios, the data preparation stage throws up major challenges.

Most data miners would like to change aspects of the data source(s) e.g. in terms of database design and data quality in order to a) reduce the level of work required and b) to improve the trends and predictions that the data mining application produces.

This research project aims to identify whether there is significant benefits to be achieved if up front data mining considerations are included in the design of the survey application. A number of areas have been identified during the research process which may have a significant impact in relation to future data mining activity. The areas are listed below and a more detail summary follows:

Database Design & Access	Survey configuration rules
Survey configuration rules	Survey configuration rules
Free form text fields	Analysis Codes
Mandatory Entries	Data entry validation rules
Data Summarisation	Maintain Historical surveys by student, by date
Ability to integrate with 3rd party systems	Maintaining a unique respondent profile

2.8.1 Database

Modern Data mining software systems work with different data sources. Generally the minimum criteria are a flat file format but they typically work with modern database management systems (DBMS). The success of the survey application is dependent on the quality of the data and results that it produces and hence the selection of the most appropriate data source is very important in terms of offering the best data repository and subsequent use

for data analysis. A modern database management system offers the most robust and structured data source for the survey application and in turn data mining.

Data preparation is a key part of any data mining process and the query facilities found within a relational DBMS is important for extracting the relevant information for data mining models. At the outset, it was envisaged that an Extraction, Transformation and Loading (ETL) process would be required prior to importing the survey data into the data mining application. After consultation with SPSS [15] (one of the leading providers of predictive analytics software), it was discovered that the ETL process can be performed within most data mining software.

2.8.2 Survey configuration rules

One of the key benefits of a survey application is that survey questions can be user defined and deployed automatically to the online survey entry module. Poorly asked questions will add little value to the information and results that can be obtained from data mining analysis. It will be difficult to guard entirely against poorly asked questions being created in the first place, however it is important that the survey application has the flexibility to reconfigure questions and quickly re-deploy the survey to respondents. This capability assumes that the timing of the re-deployment is also right to do so.

Being able to turnaround changes quickly will ensure better quality data being passed to data mining tools.

2.8.3 Free Form text fields

It is difficult to analyse datasets that contains poor data quality e.g. data errors, missing values, poorly asked questions, etc. One of the ways that organisations might minimise the impact of poor quality is providing respondents with the facility to complete free form text. For example, when a respondent has a difficulty answering a question; perhaps they don't fully understand the question or they might have two different answers. It might be very useful to provide a free form text field that allows them to qualify their answer.

Information contained in text is often referred to as unstructured data and has been historically very difficult to analyse. In general, organisations have tended to rely on structured, numerical data for their decision making. However this is changing as more and more organisations look to modern technologies that allow unstructured data to be analysed.

2.8.3.1 Un-Structured Data / Text Mining

Text Mining intelligently extracts the key concepts from text based content. The capability to understand human language is based on linguistics Natural Language Processing (NLP). Non-linguistic systems, such as statistical or probability techniques do not provide any level of understanding of the concepts and terms. While other tools based on statistical techniques simply consider a body of text as a collection of words, Text Mining intelligently extracts terms, including compound phrases. Then it automatically classifies the terms into related groups using the meaning and context of the text.

Here are just a few of the ways organisations can use unstructured data [15]:

- Understand customer preferences in detail by analysing notes fields in call center applications
- Discover common themes and important concepts in responses to open-ended survey questions
- Maximize research resources by predicting which efforts are most likely to be productive
- Monitor research trends, including the actions and relationships of colleagues in organisations doing similar research
- Predict when product components may fail or when production equipment may need maintenance, and better control both product quality and operating costs
- Predict what types of fraud, waste, and abuse are likely to occur, and where, by analysing textual information such as notes fields and e-mails
- Protect public safety and security more effectively by using predictive text analysis text to improve models of potential threats by individuals and groups

Free form text or unstructured data offers another dimension to the data that can gathered, knowing that there are facilities within data mining software that can perhaps achieve similar results to the examples outlined above.

2.8.4 Analysis codes

One of the benefits of using data mining is its ability to discover the unknown and making organisations aware of a trend that they did not know existed. In order to give data mining models every opportunity; it is worth considering how much further the data can be enriched. Typically applications are designed in a way that allows data to be analysed and grouped. For example, employees could be grouped by location, manager, grade, etc. If we could apply similar hierarchies to the survey application, the data can indeed be enriched. This can be particularly useful when using mining strategies around classifications which rely on known groupings.

There are several places where an analysis hierarchy could enhance the data mining experience. For example, it would be very useful to apply a hierarchy to the respondent profile, survey & question profiles, organisation profile, etc.

2.8.5 Mandatory entries

One of the biggest problems that data mining encounters is missing values. Respondents to a survey will typically not answer a question for various reasons – they are unsure, they don't want to commit an answer for private reasons, etc. An option to make answers mandatory should be enabled in the online survey system. The system should have the facility to configure the mandatory option by individual question as deemed appropriate by the survey administrator.

2.8.6 Data entry validation rules

Dirty data occurs for a variety of reasons. One of the ways that dirty data can be lessoned is by ensuring the appropriate level of data validation rules are applied to the online survey application. For example, lookup tables can be defined to ensure data consistency across key data fields. During data entry, field drop down boxes will ensure that users can only input pre-defined values.

2.8.7 Summarisation process

The survey application will contain some pre-processing of the data that is collected via the online survey entry module. Rather than submit the raw data that is input via the online survey system, the data should be summarised into a meaningful format that can best suited for analysis, specifically by the data mining application.

2.8.8 Maintain Historical surveys

One of the interesting topics that arose during the interview with representatives from SPSS [15] related to the idea of having the facility of mining two different surveys. With the development of a survey application that can store multiple historical surveys and indeed completed multiple times by the same individual, this surely offers a tremendous opportunity in terms of data analysis.

A word of caution was offered by Rob McCullagh [15] of SPSS in relation to attempts to compare two or more surveys as part of the same analysis. He states that it will be very important that the surveys being compared have a fair degree of commonality. This would be very useful if similar groupings are used and the surveys being compared were of similar configurations. He goes on to say that "while historical information will provide an invaluable asset, it will be difficult to compare two surveys that perhaps cover different question configurations".

2.8.9 Integration with 3rd party systems

Data mining tools are very good at identifying knowledge without any pre-conceived objectives. The more enriched the data is, the better chance the data miner has of uncovering knowledge. During the interview with representatives from SPSS [15], it was suggested that consideration should be given to incorporating a facility to allow relevant parts of the survey application to be mapped to related 3rd party systems. For example, new student aspects could be uncovered using a combination of the following student information:

- Learner preferences survey results
- Student examination results
- Student Continuous Assessment (CA) results

One of the biggest problems to achieving the combination in this scenario or for that matter many other combinations is the lack of integration between disparate systems. Sometimes there is no unique way of linking the related systems together. In the example above, the student id might be the common factor to be used, however, there may be inconsistency among different systems in terms of unique identification, etc.

To-day, major software vendors accept the fact that there will be incompatibilities between disparate systems; however to overcome this fact, facilities are provided to allow disparate systems to be integrated with their own applications. In the case of the survey application, at least one level of integration has been identified and that relates to the student or respondent.

2.8.10 Maintain a unique respondent profile

One of the benefits of an online survey application is that it affords the opportunity of holding valuable information about an individual. This is particularly important when retrieving historical survey information. Therefore, the survey application should be able to maintain a unique profile for each respondent. For example, every person who completes a survey will be assigned a unique user name and password. In addition, respondents should also be given a "forgotten user and password" option.

Over time, some respondents will complete one or more surveys and therefore, it will be important that the same unique profile, user name and password are used. When a user id and/or password are unknown, the survey application needs to be able to identify a respondent by requesting some personal information. This could include first name, last name, date of birth, mother's maiden name, email address, etc. Most web based user and password protected applications request an email address. Consideration needs to be given to the most appropriate set of questions for the target population.

2.9 Chapter Conclusion

This chapter outlines the secondary research underpinning the research objectives and demonstrates how the various areas of Survey Research, learning preferences, accessibility and data mining interrelate. This continues further with the primary research undertaken in the next chapter to determine the requirements of the proposed solution.

3. Research Analysis

3.1 Chapter Summary

The previous chapter outlined the secondary research in terms of current theory and trends in the disparate areas of generic Survey Research, learning preferences assessment, accessibility and data mining. This chapter elicits primary research in the form of structured interviews and questionnaires across the same areas. The aim of this chapter is to outline the research, discussions and findings that were carried out in pursuing the system requirements.

3.2 Introduction

The initial analysis commences with an outline of the existing system and highlights the problems associated with the current scenario. To complete the overview other commercial products and their difficulties are then analysed. The most interesting and most important task involved eliciting and understanding the needs and knowledge of the key stakeholders and subject experts.

A summary of the key issues are presented and concludes with an overview of how the analysis intends to develop the proposed system. From this study, a clear requirement was communicated, to develop an on-line learning preferences assessment tool with an integrated administration tool that will collect the data for data mining. Note that the methodology used to develop the online survey application is the object oriented approach as this is deemed the most appropriate.

3.3 Current Scenario

The current scenario involve partcipants completing a paper based screening assessment. There is at present no computerised system in place. This assessment is broken up into a number of questions for key learning areas and difficulties as presented in table 1.1. The assessment is unique and has been developed by the BUA Centre and for this reason is confidential. The completed assessment is then evaluated to determine learning style and learning difficulty (if present) and from this a broad learner profile is compiled. This process involves the manual calculation of total weightings for each learning goal and these figures are input to a spreadsheet along with respondent metadata.

Problems with the Current Scenario

Table 3.1 contains the problem areas that were identified with the current system:

Problem	Detail
Information overload	High volume of paper based data collected
Extremely difficult to analyse	Inability to extract relevant information Limited strategic applications
Difficult to see hidden knowledge	Research is restricted to statistical information Inability to forecast future best practice models
Judgment call / prone to error	Assessment may be subjective and subject to error A specialist is required to make this judgment call.
Difficult to change survey control.	Very inflexible - format and wording of Controls are difficult to change
Time consuming	Lead time between first contact and final report.
Slow reporting process	Effective retention initiatives require early detection

Table 3.1: Problems with the current system (author, 2005)

3.3 Existing Computerised Assessment Systems

BUA are not using a computerised solution but arising from the research, the following learning difficulties screening tools were identified as outlined in Table 3.2:

Screening Tool	Features
QuickScan	Generates a report highlighting the individual's learning style
Dyslexia Adult Screening Test (DAST)	Batch of assessments testing in possible areas of weakness but is restricted to delivery by trained teachers only.
Bangor Dyslexia Test	Is similar to DAST but simpler albeit requires a trainer for delivery
Lucid Adult Dyslexia Screening LADS	Tests phonological processing, working memory and lexical access.

Table 3.2: Existing computerised assessment tools (author, 2005)

Problems with Computerised Assessment Systems

There are many difficulties with the existing systems:

- Results lack specific practical recommendations, plans and outcomes
- Limited analysis & evaluation
- Many computerised assessment tools are not accessible to many people with disabilities.
- None of the existing computerised assessments look across all the specific learning difficulties and can be self-administered and the response tailored to the individual.

3.4 Research Interviews

In order to develop a comprehensive set of user requirements, a number of interviews were conducted as outlined in table 3.3. This extensive interviewing ensured comprehensive information gathering for the system requirements from the four disparate project areas; Generic Survey Research , Learning Preferences Assessment, Accessibility and Data Mining. (See Appendix F for sample research questionnaire).

What emerged was the variety of interests and concerns for the various stakeholders -students, administrators, college decision makers, academics, learning support workers and data miners. The challenge being how best to accommodate all interests.

Interviewees	Role	Context
Larry McNutt	Head of School of Informatics & Engineering, ITB	Data Mining
Diarmuid O'Callaghan	Registrar, ITB	Data Mining
Amanda Kirby	Specific Learning Styles Subject Expert, Dyscovery, Wales	Learning Styles / Difficulties Data Mining
Dawn Duffin Suzanne McCarthy	BUA Manager BUA Psychologist	Learning Styles / Difficulties
Geraldine Gray Orla McMahon Colm McGuiness	Academics ITB	Learning Styles Data Mining
Susan Bergin	Academics, NUI, Maynooth	Survey Research
ITB Students	Focus group and first year students	Learning Styles
Rob McCullagh	Data Mining Consultant, Habaca/SPSS	Data Mining
Dr. Maura Reagan	Statistics & Data Mining Lecturer, Trinity College	Data Mining
Dr Chris Singleton	Dyslexia at Third Level Expert	Learning Styles Accessibility
Joshue O'Connor	National Council for the Blind in Ireland	Accessibility

Interviewees	Role	Context
	Centre for Inclusive Technology	
Sally Fincher	Head of Computers and Research Group Kent University	Learning Styles
Suzanne Walsh	Assistive Technology Centre ITB	Accessibility
Denise Jackson	Jackson Assistive Technologies	Accessibility

Table 3.3: Survey Research Tool Interviews (author, 2005)

3.5 Generic Survey Research

As outlined in section 2.3.3, a collaborative partner assisted in identifying the issues involved in designing a generic Survey Research. Susan Bergin, a PhD Researcher at the Department of Computer Science, NUI Maynooth (Supervisor Professor Ronan Reilly) collaborated with testing the generic suitability of the design and implementation of the survey tool. As part of her research Susan required survey data. The target population was first year students studying software development on the computer science course at the Institute of Technology, Blanchardstown academic year 2004/2005.

The following instruments were identified as possible surveys:

- Identifying factors that influenced performance for first year programmers
- Motivation
- Learning Strategies

3.6 Learning Preferences Assessment

Management at the ITB highlighted the importance of gaining a complete view of the student; the ability to merge *existing* datasets such as Examination results, Attendance details, Course Board minutes & actions.

Student Retention is a core issue for management and they are interested in strategies which enable the individual to be more responsible for their own learning.

Students are primarily interested in how to make their own learning experience more effective. Students, generally, are not interested in collated results; how the student body is performing, retention initiatives or changes in teaching methodologies which do not have a direct impact on their progression.

Academics are interested in both individual learning profiles for students and overall collated results resulting in adopting best practice teaching methodologies in order to have most effective impact on facilitating student learning.

BUA recommend a very pragmatic approach to the content of the individual learning plan addressing strengths and weakness to allow the student reach their full potential. In other words the more specific the recommendations, the more possible impact for the student.

There is evidence to suggest that there is considerable overlap between each of the specific learning difficulties (outlined in table 1.1) [7]. Many assessment tools target a specific learning difficulty. A coordinated learning plan is required to address their individual specific difficulties (not multiple separate plans that address the disability label).

As a result of the research discussions with the collaborative partners a learning preference model (as shown in figure 3.1) has been proposed to measure learning styles and difficulties. It is intended that this model will be incorporated into the design and development of the survey application. Essentially, the methodology applies a weighting to each question within a learning area or goal. The accumulated weighting measure can determine a learning style.

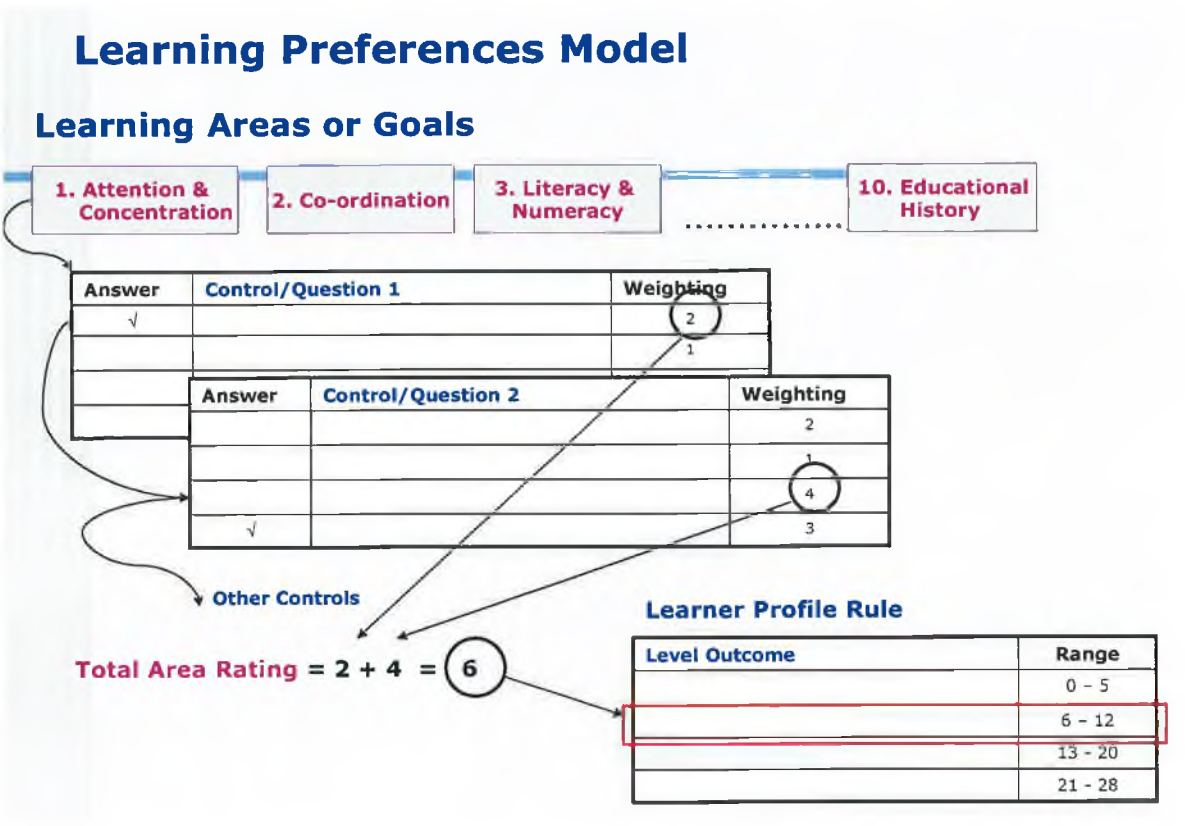


Figure 3.1: Learning Preferences Model (author, 2005)

3.7 Accessibility

One of the greatest barriers to the deaf student's education is the accessing of course texts and research papers. A deaf person whose preferred language is sign language will not necessarily have acquired fluency in accessing written information in English [17]. Survey accessibility for deaf students requires the survey content to be written in a manner that facilitates deaf students.

The need to gain more insight into the impact of learning styles to aid policy making was frequently mentioned. If we had additional statistics, additional funding could be sought to address specific needs.

National Centre for Inclusive Technology

The National Centre for Inclusive Technology, based in the National Council for the Blind provide consultancy in Inclusive Design and independent verification of accessibility through auditing and testing services. They stress the importance of identifying the goals of accesibility as the initial starting point for any accesibility project. The common pitfalls are attempting to provide accessibility for a particular item of Assistive Technology e.g. JAWS speech for blind users. The best practice approach is to conform to HTML coding standards independent of these assistive technology user agents, rather than conforming to proprietary systems.

The importance of user testing cannot be over estimated and CFIT flag the pitfall of only testing using HTML validators such as Bobby (see Appendix E).

The development of an accessibility statement actively encourages user participation, creating a continual testing process, by encouraging users to highlight difficulties, provide suggestions and feedback to continually improve the accesssible feedback and also incorporate new facilities and features.

Innovatively CFIT use multimedia capture during web accessibility testing to capture the user experience in real time. This visual record powerfully shows the level of frustration when a user attempts to read web content with speech systems when the WCAG guidelines are not adhered to.

Colour contrast provides accessibility for certain visual impairments. The facility to 'switch styles' allows the user to easily change the user interface; colour schemes, font style and size etc. For many users this is a preference but for some it can be the difference between inclusion or exclusion. Additionally the environment also needs careful consideration, background lighting, glare, natural lighting etc are examples which may make a difference.

3.8 Data Mining

Before working with Data mining, it is very important to explore what the organisation expects to gain from data mining. Many key people were involved in these discussions. A summary of their input is outlined below:

Data Mining Experts

Data miners were in agreement over the high volume of data mining projects that fail. They stressed the importance of following an appropriate data mining methodology, developing data warehouse type data structures and the importance of capturing the correct data to gain an accurate view of learner profiles. The data mining methodology must always commence with a thorough understanding of the nature of the business and all nuances involved.

Management

Student Retention is a core issue for management at ITB, whom indicated their interest in gaining information about students who are not successful i.e. those students who either leave early or are unsuccessful in examinations.

BUA

Representatives from BUA are keen to elicit new or unknown information about their data. It was also high on the BUA agenda that data mining could perhaps lead to some new knowledge on secondary disability. Frequently students may present themselves with a particular disability, for example, visual impairment but actually show signs of a secondary disability, for example, dyslexia. Another potential opportunity for data mining might include new knowledge about secondary disabilities which often gets ignored.

3.8.1 Un-Structured Data/ Text Mining

Instances exist where the richness of data gathered in Survey Research is diminished by use of structured questions. The exclusive use of tick boxes prevents the free flow of possible 'nuggets of information' as can be entered in comment text boxes.

Incorporate Text mining into the Survey Application

Data richness can be derived from free format text fields on survey forms allowing the user to include comments.

Text mining and Data mining

For each comment field an index of concepts, as well as the frequency and class of each concept is returned. This distilled informed can be combined with other data sources and used with traditional data mining techniques such as clustering, classification and predictive modeling . Questions to explore include which concepts occur together? What else are they linked to? What do they predict? How do concepts predict behaviour? Data mining and text mining can be combined to deliver greater results than is available from either structured or unstructured data alone.

3.9 System Users

There are many users of the proposed framework, incorporating the survey application and data mining. Each of the roles listed in table 3.4 will have a core level of interest namely strategic, operational and/or personal. Table 3.4 contains a matrix of how different roles are expected to interact with different areas of the framework.

Role	Individual Learning Preferences	Collated Data Mining Results
Head of School		Y
Head of Dept		Y
Registrar		Y
Psychologist	Y	Y
BUA Manager	Y	Y
Academics	Y	Y
Students	Y	

Table 3.4: Interests of Key Stakeholders (author, 2005)

3.10 Proposed System

A number of technologies have been identified during the research phase and now form part of the proposed framework for the generic survey application in conjunction with the Learning Preferences requirements. The technology framework will consist of the following three components:

1. Survey Administration
2. Online survey entry module
3. Data Mining Analysis

1. Survey Administration

It is expected that this module will be used by psychologists to design surveys. The main features will include:

- Ability to configure and maintain the survey questions and specific assessment weightings
- Facility to configure and maintain organisational details and student details
- Generate individual student's results for data analysis
- Generate an individual's Learner Profile report

2. Online Survey entry module

It is expected that this module will be used by students to complete survey responses online. The main features will include:

- Automatic deployment of survey questions
- Include accessibility features for visually impaired, hearing difficulties, etc
- Students can define their own user settings.

3. Data Mining Analysis

- It is expected that this facility will be used by psychologists and organisation decision makers to assess survey results.

3.11 Chapter conclusion

The current paper based system has numerous difficulties. Existing computerised systems tend to address a specific disability, are not configurable and do not extend to data mining. The next challenging task of developing the system requirements was greatly assisted by the interviews and completed questionnaires. The requirements draw together the generic Survey Research, learning preferences, accessibility, data mining and text mining. These integrated requirements will become the focus for the optimal design focus in the next chapter.

4. Design

4.1 Chapter Summary

This chapter describes the technical approach and design that will apply to this research project. A key part of this research is to explore and deliver technology that can advance the information that can be gleaned from survey based systems in conjunction with data mining. Figure 4.1 illustrates an overview of the technology areas that have been investigated which now form the basis for the proposed development. Note that it is not intended to deliver a fully functional solution, however it is intended that the design of the proposed solution should consider all of the relevant aspects of the research subject. From the design, a strategic prototype that incorporates all of the chosen technologies will be developed which has the objective of providing conclusive evidence with regard to the project's research objectives.

A number of tools and technologies were considered to support the solution architecture. For each of the technologies, the chapter incorporates the following questions:

a) What were the key criteria that were used to select the appropriate technology?

b) What was the reason for choosing a technology over others?

c) Did the chosen technology deliver conclusive evidence in relation to the research subject area?

4.2 Solution Architecture

The solution architecture will comprise three primary technologies as shown in figure 4.1 and these include the following:

1. A Database Management system that will be used to design a robust secure database
2. A web based Survey System will be developed using modern scripting tools
3. A Data Mining tool that can interact with the data collected from the web based survey system.

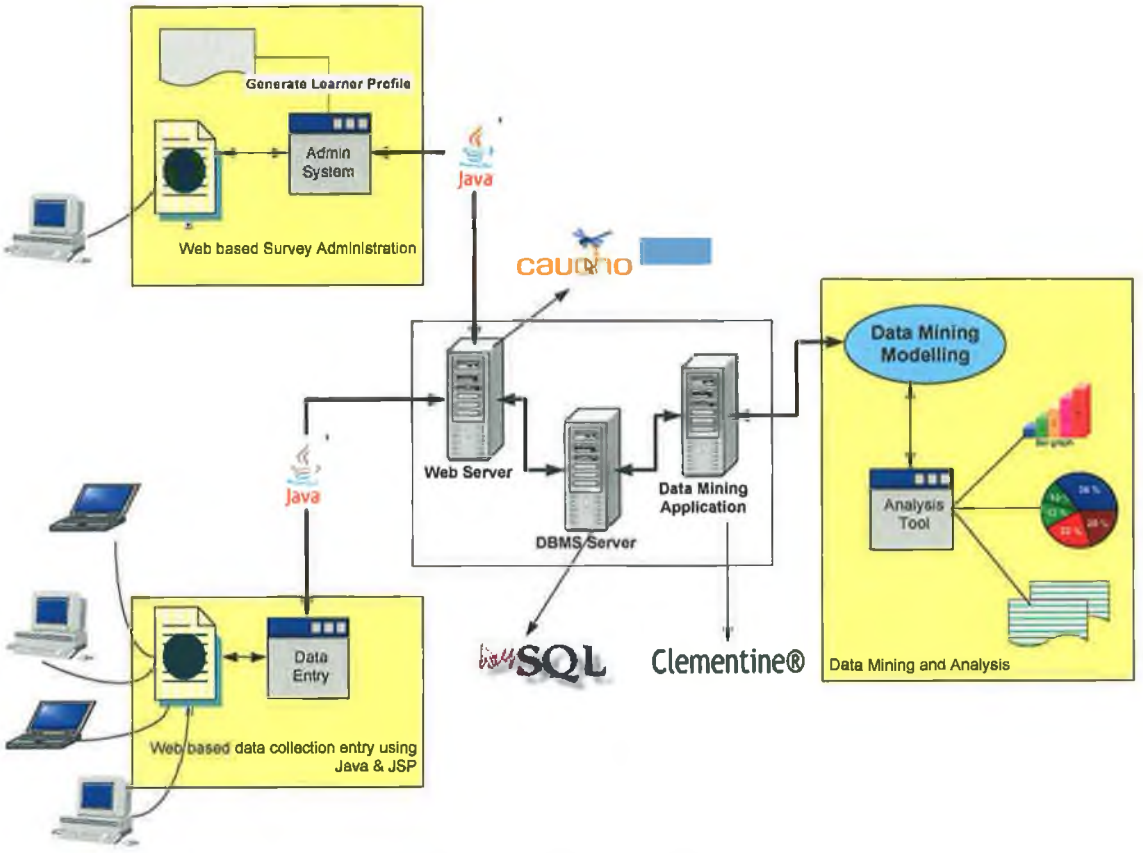


Figure 4.1: Proposed Solution Architecture (author, 2005)

4.3 Technology

This section covers the following technologies:

- Database Management System
- Web Server
- Web Development Tools
- Data Mining software

4.3.1 Database Management System

Initially Microsoft Access was chosen as the database but after subsequent research, the following limitations deemed it not suitable [18] [19]:

- Web Responses must be imported (not "real-time")
- Changes in user logins, relationships etc. must be uploaded to Web Site

- Security is weak
- Generally used as a personal or single-user application

The initial database design was performed using MS Access but after further investigation, MySQL was chosen as the DBMS to support the survey application. MySQL is an open source database management system that is gaining a lot of credibility among the academic and business community [18] [19]. Some of the reasons for choosing MySQL include the following

- Data is available in "real time" for reports and enquiries
- All survey changes (e.g. adding logins) are in real-time
- All data is stored in a single database
- It is free of charge and therefore no licensing issues

4.3.2 Web Server

The criteria for choosing a web server for this research project included the following;

- What is the platform that the Web server be deployed? For example, the chosen platform is Windows.
- What types of technologies need to be supported? Many Web servers are better at serving static content than others but perform poorly when required to serve content generated by JSP pages or server-side Java technology-based applications.
- The Web server must work alongside the chosen server-side language
- The chosen web server should have an acceptable technical support to deploy and maintain the Web server
- What Connectivity Tools are available in conjunction with Database Management Systems (DBMS)? For example some web servers contain in-built connectivity capabilities.
- The web server should ideally be available as open source and free with low cost implications

Two web servers were evaluated during the research phase; Apache Tomcat and Resin from Caucho Technology. While both products satisfied all of the criteria outlined above, Resin was chosen as the web server primarily on the basis that there was a local support group available.

Resin Web Server

The Resin web server as per figure 4.2 will be used to distribute the survey application to students and administrators using a web browser. Resin provides a fast standalone web server; for many sites, the standalone web server is ideal because of its performance and because it is easier to configure and maintain than using a separate web server [20].

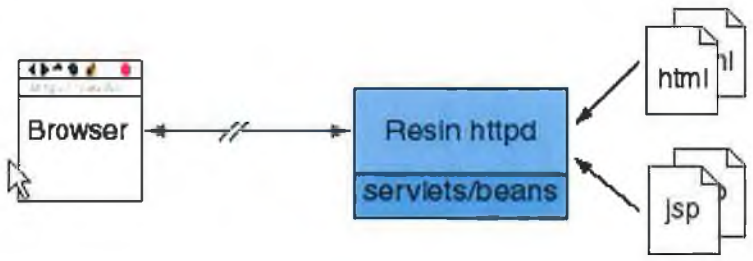


Figure 4.2: Resin Web Server (Source: Caucho)

Resin supports the use of any database that is available with a JDBC driver which includes MySQL database.

4.3.3 Web based Development Tools

Web development tools will provide the user interface and database interoperability. Hyper Text Markup Language (HTML) is the language of the Web. Using HTML on its own is not enough to provide the functionality required by the survey application and for this reason, client and server-side scripting languages are required. Here are a few examples of features that the application should provide:

- Make it easier to edit contents of the Web page, by updating the survey application database rather than the HTML code itself.

- Create pages that will be customised to display only content that will be of interest to a particular user (e.g. survey questions).
- Retrieve, display and update survey details contained in the Web page.

A Web authoring tool (Dreamweaver from Macromedia) was considered which allows Web developers to generate HTML and JavaScript source code while viewing the site as they develop. However it has been decided to use the standard Textpad facility to develop the code as it was recommended by CFIT that developing in code enables stronger accessibility control.

4.3.3.1 Scripting Languages

Client Side Scripting

Client side scripting refers to scripts that are executed by the client which is the browser. Client-side scripting enables interaction within a webpage. The code required to process user-input is downloaded and compiled by the browser. Client-side scripting languages include JavaScript and VBScript.

Javascript has been chosen as the client side scripting Language for the web component of ths research project. It is used to create a sophisticated and interactive front end that can be executed by the browser itself and does not require data to be transmitted to the server.

Server-side scripting

Server-side language is scripting language which runs on the server and not on the user's browser machine. With server-side scripting, completing an activity involves sending information to another computer (or server) across the internet. The server then runs a program that processes the information and returns the results, typically a webpage.

Figure 4.3 shows an example of how a server side language such as JSP operates alongside a web browser, web server and database management system.

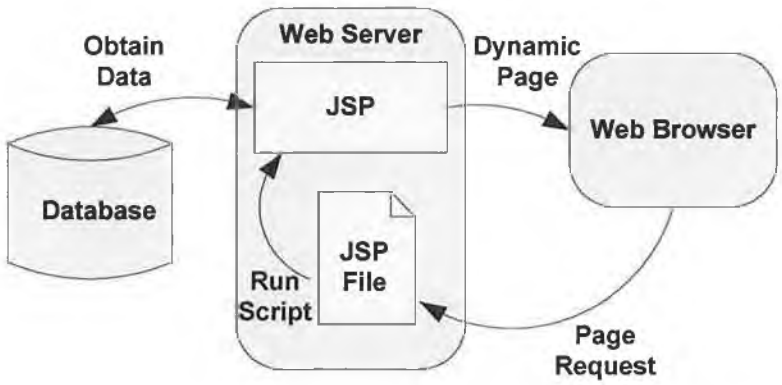


Figure 4.3: Interpreting JSP (author, 2005)

Server-side scripting languages include PHP, ASP, JSP, Coldfusion and PHP. Appendix I contain a comparison of the server-side scripting langauges that were researched for this project.

Reasons for Choosing JSP

JSP was chosen as the preferred server side scripting tool. The reasons are as follows:

- JSP is one of the most widely used server-side scripting languages
- JSP and Java code is typically free.
- Java and JSP are platform independent.

A sample JSP program is included in Appendix J.

4.2.3.2 JSP and Java Beans

JSP offers many features for easy and quick development of Web applications. However, if such features are used without some planning and structure, the JSP code can become a mix of various HTML tags, JSP tags, and Java code that is difficult to follow, debug, and maintain.

The objective is to have JSP code that resembles HTML to the greatest possible degree, by moving all processing code into Java Beans [21]. The development of the web based components for this research project includes JSP and Java Beans.

What is a Java Bean?

A Java Bean can be defined as a reusable software component. What this means is that it is possible to write a Java Bean that can then be used in a variety of other Java based softwares such as applications, Servlets or JSP pages. In this way we can define the business logic within a Java Bean and then consistently use that logic in separate applications. [22]

Benefits of Java Beans

We now have three ways of writing code to be used by a JSP. These are,

1. Place the code at the start of a JSP in a declaration,
2. Use an include statement to reference another file which contains the code and now
3. Package the code in a Java Bean

The method to use and when depends, as always on the specific circumstances. Writing the code within a JSP page is certainly the most straightforward but it does limit the amount of code reuse. Using include statements does lend itself to a certain amount of code reuse and certainly it is the only way to allow reuse of chunks of HTML.

By using Java Beans it is possible to fully separate the business logic from the generation of the display. This is an important philosophy that leads to better structured and more maintainable systems. In most cases a JavaServer Page is used to dynamically generate display and also to handle the user interaction. The Java Bean would take over when some complex data processing needs to be performed or when databases or the file system needs to be accessed [22].

A sample Java Bean is included in Appendix K.

4.2.3.3 JSP and DB Wings

During software development a very useful tool called DB-Wings was used to automatically create the JSP code for table maintenance. DB-Wings is a utility developed by an ITB colleague Matt Smith and his associate Sinead Murphy and who granted permission for its use in this research project. Its main features include the following:

- Create front-end JSP applications for local and remote databases
- Automatically generates Java Beans and JSP which can be modified for further enhancement
- Future enhancements include interface configuration features and security options
- Currently only caters for a single field in a primary key which as a result requires further code amendment.

See appendix H for technical details.

4.3.3.4 JSP and Accessibility

As per the National Centre on Accessible Information Technology in Education at the University of Washington [23] who state that Server-side scripting does not, in and of itself, present accessibility problems. Like all web content, however, server-side scripts must produce content that follows principles of accessible design. The Web Content Accessibility Guidelines 1.0, which are defined by the World Wide Web Consortium (W3C), include four guidelines that specifically address scripting issues:

> ***6.3*** *Ensure that pages are usable when scripts, applets, or other programmatic objects are turned off or not supported. If this is not possible, provide equivalent information on an alternative accessible page. For example, ensure that links that trigger scripts work when scripts are turned off or not supported (e.g., do not use "JavaScript:" as the link target). If it is not possible to make the page usable without scripts, provide a text equivalent with the NOSCRIPT element, or use a server-side script instead of a client-side script, or provide an alternative accessible page as per checkpoint 11.4.*
>
> ***6.4*** *For scripts and applets, ensure that event handlers are input device-independent.*
>
> **6.5** *Ensure that dynamic content is accessible or provide an alternative presentation or page. For example, in HTML, use NOFRAMES at the end of each frameset. For some applications, server-side scripts may be more accessible than client-side scripts.*
>
> **9.3** *For scripts, specify logical event handlers rather than device-dependent event handlers.*

4.3.4 Data mining software

The criteria that were used for choosing a data mining tool that can support Survey Research contained the following considerations:

- What track record does the data mining tool vendor have in relation to Survey Research
- Organisation's data generally originates in a number of heterogeneous databases in different formats. To avoid unnecessary "interfacing" work, the data mining tool should be able to read their various data formats e.g. MS Excel, ODBC, etc
- It is thus critical to ensure that the tool selected is able to assist effectively, in data preparation
- Ensure that the data mining tool implements modeling algorithms that meet the needs of Survey Research. Ensure that the business objective can map to an adequate model type.
- Must be able to deal with a variable size of data sets
- The User Interface features should contain minimum criteria in terms of:
  - Should have a modern Graphical Layout:
  - Support Drag&Drop/Visual Programming:
  - Should have On-line Help to support all types of users
- Can the vendor provide the software to the research project for free?

A comparison of three data mining software products that were reviewed is shown in table 4.1.

Data Mining Tool	Features
Clementine http://www.spss.com/	Regarded as the leading data mining benchmark
	Supports the entire business process of data mining
	Designed around the industry standard CRISP-DM methodology
	Focuses on solving specific organisational problems
	User friendly interface
	Multiple successful references
SAS http://www.sas.com	Provides user friendly front end for users
	Designed based on the SEMMA (Sample, Explore, Modify, Model, Access) methodology
	Allows the user to avail of many different algorithms
	Perceived to be harder to use than Clementine
	Used over Clementine when the amount of data to be mined is significantly higher
Oracle 10g Data Mining	Oracle Data Mining is an option to Oracle 10g Database

Data Mining Tool	Features
http://www.oracle.com	Enterprise Edition
	Supports a wide range of algorithms
	Perception that an implementation requires technical support
	Targeted at organisations that are committed to the Oracle platform as data mining is embedded in the Oracle database.

Table 4.1: Key features of data mining tools investigated (author, 2005)

Reasons for choosing Clementine

Clementine was chosen as the preferred data mining tool. The reasons for this choice are outlined below:

- ITB have a relationship with SPSS, Inc, the owners of the Clementine product which included a software licence that can be used as part of this research project.
- As a result of this relationship, SPSS were very agreeable to provide advice on aspects of the Clementine solution. This included several phone calls and meetings.
- Clementine is one of the leading data mining software packages on the market and was indeed very attractive. In a poll conducted by KDNuggets in 2005 [24] comprising 860 votes and titled "Data Mining Tools used in 2005", Clementine clearly came out on top as shown in Appendix O.
- There were a number of references among Educational authorities that were using Clementine for key statistical analysis
- SPSS appeared to have a focused strategy to help organisations get the best out of their Survey Research data.

4.4 Survey Application

A survey application prototype will be designed and developed using the technologies outlined in figure 4.1. The survey application will contain two web based modules:

- Administration module
- Online survey entry module.

The administration module will be primarily used by a survey administrator to perform the following tasks:

1. Setup any classifications that relate to Organisations, Students, Surveys, Survey Questions and Accessibility.
2. Define the organisation (who is running the survey)
3. Setup the survey
4. Define the survey questions
5. Generate Survey Results and Learner Profiles

The online survey entry module is a web based application that students will use to enter a survey. The application will consist of the following process:

1. The student will launch the online survey using the appropriate URL
2. A security page will be presented that will require the input of a valid user name and password. If the student doesn't have one, they will be able to register and add student profile details.
3. When a valid user name and password is accepted the student will be presented with the relevant survey questions and instructions. This will allow the student to commence the survey.

4.4.1 UML Design

The heart of object-oriented problem solving is the construction of a model. The model abstracts the essential details of the underlying problem from its usually complicated real world. Several modeling tools are wrapped under the heading of the UML™, which stands for Unified Modeling Language™ [25].

The UML is a family of graphical notations, that help in describing and designing software systems, specifically developed using object oriented approaches. At the centre of the UML are its nine kinds of modeling diagrams. For the purpose of this research project, two kinds of diagrams have been used; Use Case and Class diagrams.

Use case diagrams describe what a system does from the standpoint of an external observer. The emphasis is on what a system does rather than how. Use case diagrams are closely connected to scenarios. A scenario is an example of what happens when someone interacts with the system [25].

A **use case** is a summary of scenarios for a single task or goal. An actor is who or what initiates the events involved in that task. Actors are simply roles that people or objects play [25].

Use case diagrams are helpful in three areas.

- Determining features (requirements). New use cases often generate new requirements as the system is analysed and the design takes shape.
- Communicating with clients. Their notational simplicity makes use case diagrams a good way for developers to communicate with clients.
- Generating test cases. The collection of scenarios for a use case may suggest a suite of test cases for those scenarios.

A high level Use Case diagram has been prepared for the survey application and is shown in figure 4.4. A separate Use Case diagram exists for each of the scenarios shown. Figure 4.5 contains the Manage Survey Data Use Case diagram. Figure 4.7 and figure 4.8 outline the Enter Survey Use Case and the Analyse Results Data Use Case diagrams.

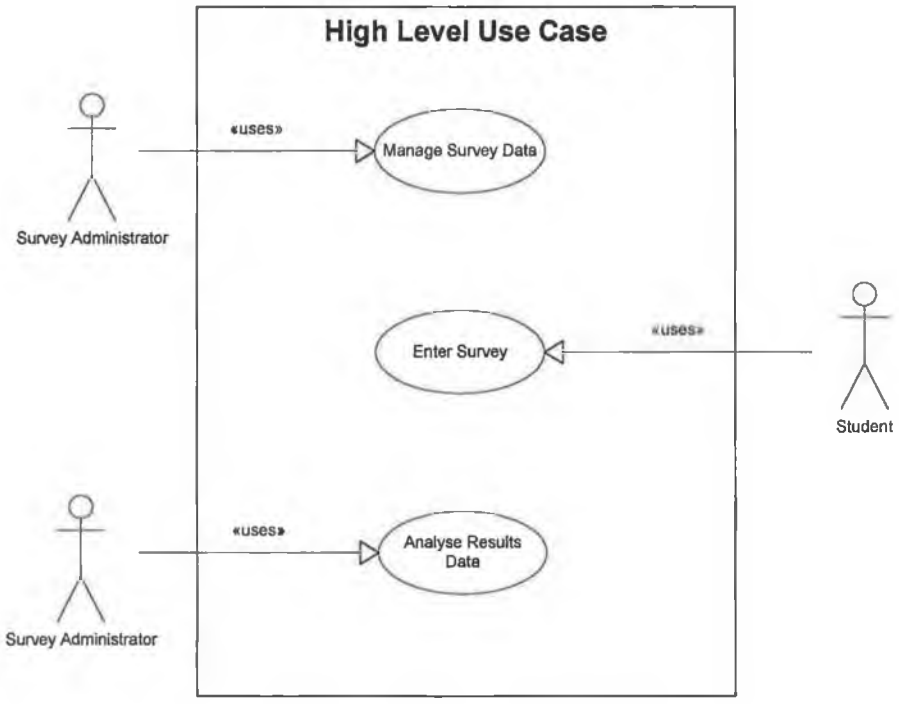


Figure 4.4: High Level Use Case diagram (author, 2005)

The Manage Survey Data Use Case diagram in figure 4.5 demonstrates the level of scenarios in relation to the survey administration module.

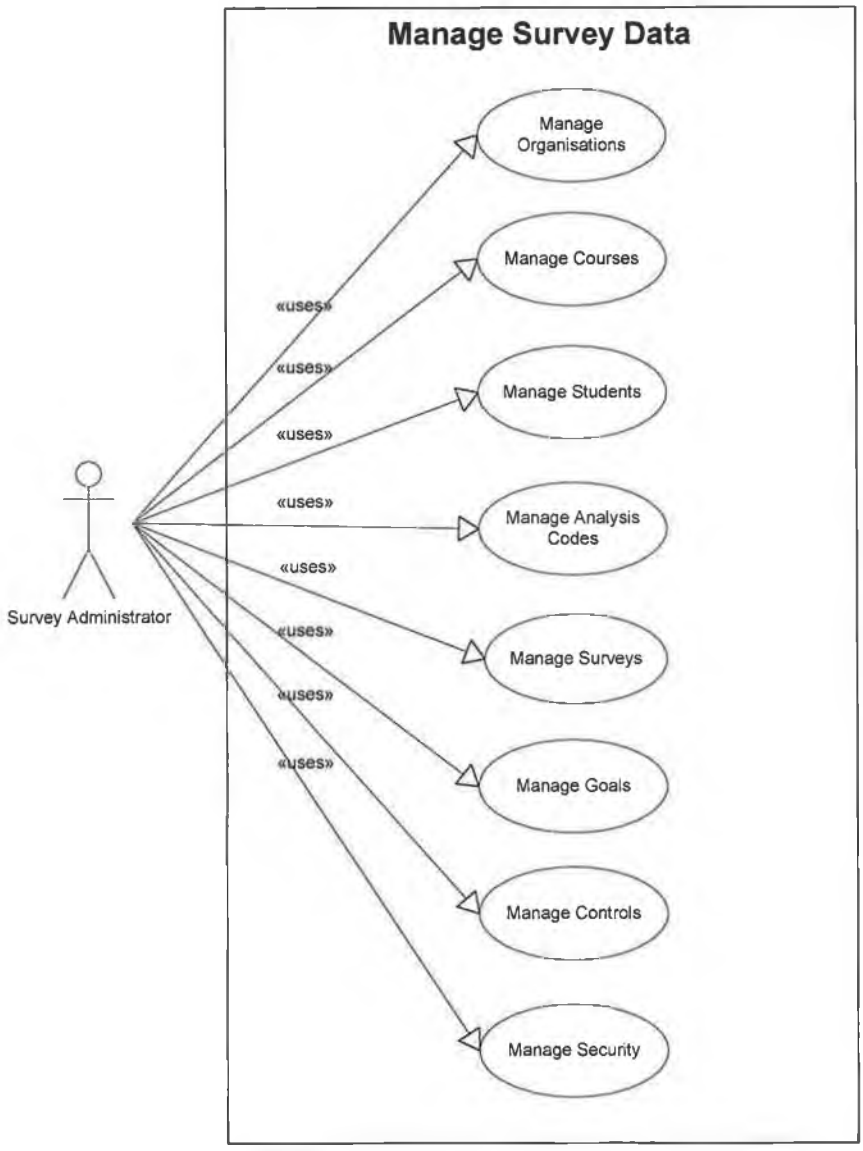


Figure 4.5: Manage Survey Data Use Case diagram (author, 2005)

Following on from the Manage Survey Data Use Case diagram, further Use Case diagrams are defined. Essentially a separate Use Case diagram will exist for each of the oval shaped use cases. For example, the Manage Surveys Use Case diagram is outlined in figure 4.6.

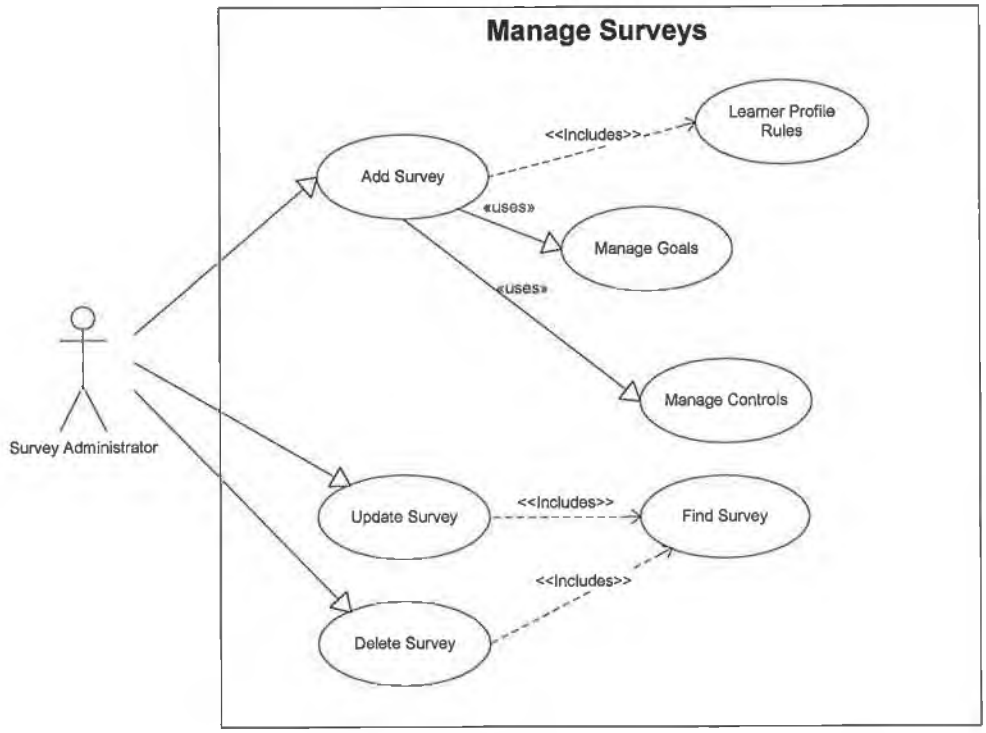


Figure 4.6: Manage Surveys Use Case diagram (author, 2005)

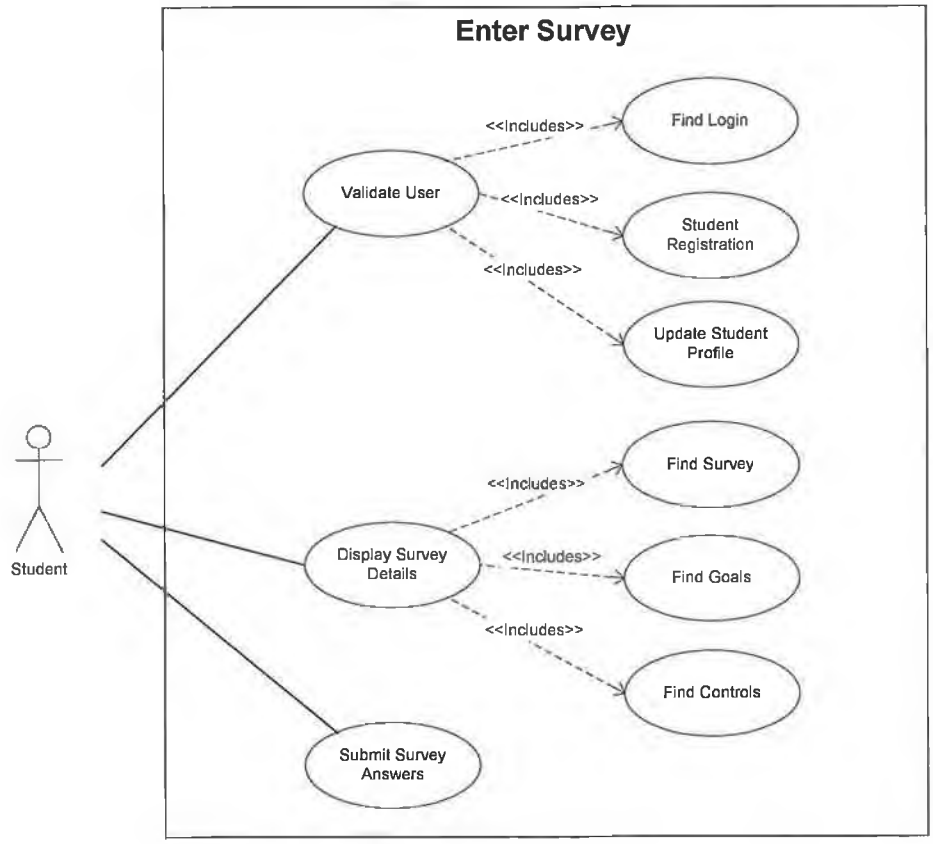


Figure 4.7: Enter Survey Use Case diagram (author, 2005)

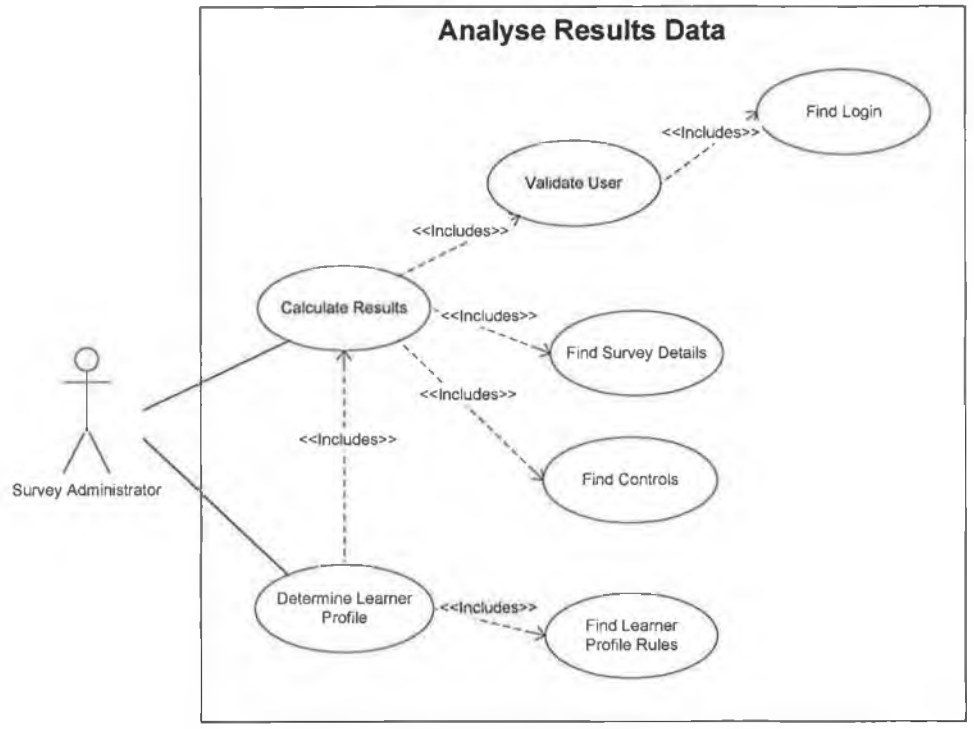


Figure 4.8: Analyse Results Data Use Case diagram (author, 2005)

A **Class diagram** gives an overview of a system by showing its classes and the relationships among them. Class diagrams are static; they display what interacts but not what happens when they do interact. UML class notation is a rectangle divided into three parts: class name, attributes, and operations. The class diagram in figure 4.9 has two kinds of relationships [25].

Association - a relationship between instances of the two classes. There is an association between two classes if an instance of one class must know about the other in order to perform its work.

Aggregation - an association in which one class belongs to a collection. An aggregation has a diamond end pointing to the part containing the whole. In figure 4.9, Client_Survey_Header has a collection of Client_Survey_Details.

In this example at figure 4.9, the classes represent the proposed tables and attributes are defined by the fields within each table. Note, an example of an operation is shown under the Client_Survey_Details class.

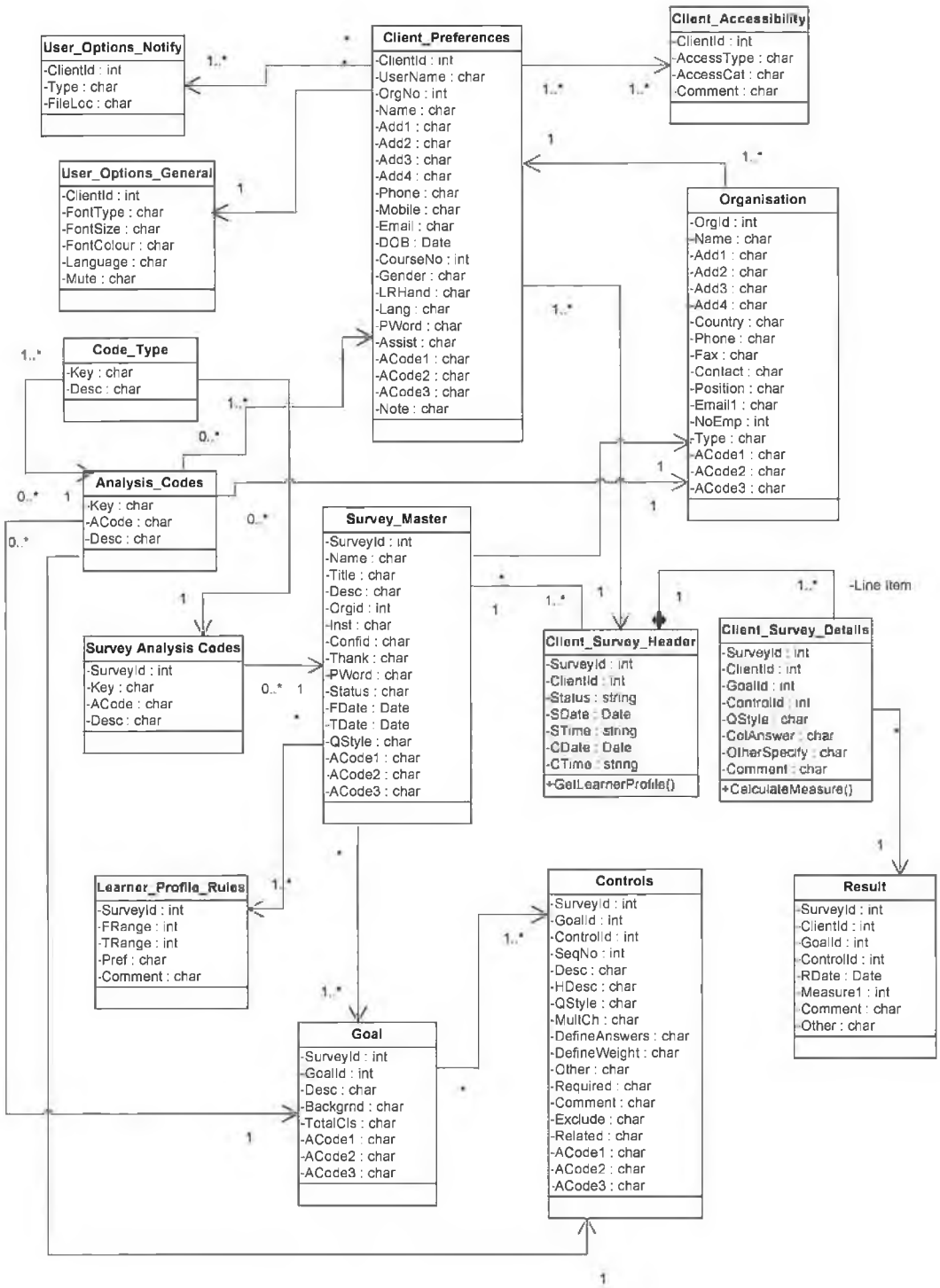


Figure 4.9: Survey Application Class diagram (author, 2005)

An association has two ends. An end may have a role name to clarify the nature of the association. For example, a Client_Survey_Details is a *line item* of each Order.

A navigability arrow on an association shows which direction the association can be traversed or queried. A Client_Survey_Details can be queried about its Item, but not the other way around. The arrow also lets indicates who "owns" the association's implementation; in this case, Client_Survey_Details has an Item. Associations with no navigability arrows are bi-directional.

The multiplicity of an association end is the number of possible instances of the class associated with a single instance of the other end. Multiplicities are single numbers or ranges of numbers. In our example, there can be only one Organisation for each Survey, but an Organisation can have any number of Surveys.

4.4.2 Database Design

There was no existing database to use as the basis for the new database design and therefore the database design had to start from scratch. This section outlines a summary of the tables that have been incorporated into the new database design.

Student / Organisation / Course:

For each organisation a number of courses may exist. The course details must be recorded. The student details must be stored, linking the student with a course and an organisation.

Analysis Codes

Many tables contains one or more analysis code fields which can be used for statistical analysis and reporting. A separate analysis code table exists to ensure that valid entries are input. Each analysis code contains a unique code type field and description. The code type is very important and is used to apply the appropriate validation to the specific analysis code field. When a user enters a value into an analysis code field, a drop down list will contain only the valid entries associated with the code type e.g. Hearing Type, Organisation

Type, etc. Code types are defined in a separate table. The following code types have been predefined as per table 4.2:

Code Type	Table	Description
CASUBC	User Accessibility	Accessibility Category
CLAN01	Control	Client Analysis 1
CLAN02	Control	Client Analysis 2
CLAN03	Control	Client Analysis 3
CSAN01	Course	Course Analysis 1
CSAN02	Course	Course Analysis 2
CSAN03	Course	Course Analysis 3
CTAN01	User	Course Analysis 1
CTAN02	User	Course Analysis 2
CTAN03	User	Course Analysis 3
CLGENDER	User	Gender
CLHAND	User	Left or Right Handed
GLAN01	Goal	Goal Analysis 1
GLAN02	Goal	Goal Analysis 2
GLAN03	Goal	Goal Analysis 3
ORAN01	Organisation	Organisation Analysis 1
ORAN02	Organisation	Organisation Analysis 2
ORAN03	Organisation	Organisation Analysis 3
ORCTRY	Organisation	Country
ORTYPE	Organisation	Type
SVAN01	Survey Master	Survey Analysis 1
SVAN02	Survey Master	Survey Analysis 2
SVAN03	Survey Master	Survey Analysis 3

Table 4.2: Pre-defined Code Types (author, 2005)

In the case of the student accessibility table, analysis codes exist for Accessibility Type, and Accessibility sub-category. An example of valid entries is included for the accessibility sub-category in table 4.3:

Value	Description
000	Unknown
002	Myopia
003	Short Sighted
004	Long Sighted
005	Stigmatism
006	Used to wear glasses
020	Deaf

Table 4.3: Accessibility sub-category valid entries (author, 2005)

It is not possible for the system to second guess what every organisation might require for analytical purposes. Therefore a number of user defined analysis codes have been incorporated into several tables. Up to three additional analysis codes have been provided in the following tables:

Organisation table	Client table
Course table	Goal table

Essentially, an organisation can define their own code type description and unlimited number of three character entries per analysis code.

Student Information

Before a student can complete a survey, he/she must register their personal information. This information will be useful for basic contact details but in addition pertinent personal details will be useful when combined with survey information during data mining. The key elements of the student table include the following:

Student Id	User Name	Name, Address, Phone No, Mobile No, Email
Date of Birth	Gender	Right or Left Handed
Assistance required to complete survey	Analysis Codes	

SPSS [15] made a very useful recommendation around authenticity of students in relation to keeping a track of historical surveys. It would indeed be very useful to maintain a history over similar or different surveys. A major challenge is to ensure that duplicate student references are not created in the event that a respondent does not remember their previous reference / ID, etc. The database design and the subsequent online screen validation of student registration should ensure uniqueness. A number of options need considering which include using the following combinations of fields; Name, Date of Birth, email address. It also worth exploring what is available by way of fuzzy logic facilities that will ensure consistency and increase the quality of the information.

Student Accessibility

Some students who will complete the online survey will have more than one accessibility issue. To ensure that all of the information is recorded, respondents will be able to declare multiple accessibility types. An accessibility type in this context of the Learning Preferences covers the broad areas of Hearing, Blindness and Impaired Vision. A further sub-category will allow an additional classification. The key elements of the student accessibility table include the following:

Student Id	Accessibility Type	Sub Category
Relevant Comment		

Survey

An entry must be created in the survey master table before any survey data collection can commence. The key elements of the survey include the following:

ID, Name,	Title	Instructions,
Confidentiality statement,	Thank You statement	Survey Status e.g. active, inactive
Effective dates	Default Question Style	Survey Analysis Codes

Goals

Goals or also known as sections are user defined by survey id. A goal in the context of learning preferences could be a specific learning area or classification. The following are typical examples of goals:

- Literacy & Numeracy
- Concentration
- Communications
- Coordination
- Learning Styles: Visual, Audio, Kinaesthetic

The key elements of the Goal definition include:

Goal Id, Survey Id,	Description,	Background information
Goal analysis codes		

Controls

Controls or also known as questions are defined uniquely by Survey Id, Goal Id and Control Id. A goal is made up of a number of controls relating to that goal only. Here is an example of how a control or question might be presented:

	Hardly ever	Occasionally	Sometimes	Most of the time
Do you get confused when given several instructions at once?	○	○	○	○

The key elements of the Control definition include:

Question Style	Multiple choice answers	Answer Options
Weightings	Other – Please Specify	Allow comment
Answer Required	Control analysis codes	Exclude from Results
Exclude from Results		

For accessibility, a facility will be provided to allow an audio file or .wav file to a control. If required, a student with visually impaired difficulties will be able to play an audio of the control during the online survey.

Student Survey Responses

A header and detail table will be maintained when a student completes a survey. The key elements of the survey header table include the following:

Survey Id	Student Id	Status
Start Date, Completed Date	Start Time, Completed Time	

The key elements of the survey detail table include the following:

Survey Id	Student Id	Goal Id
Control Id	Question Style	Text Answer
Column Answer	Other – Please specify	Comment

Results

When all of the surveys have been completed, a specific process in the survey application will generate a set of results or measures for each combination of Survey Id, Student Id, Goal Id and Control Id. The key input to this process will be the respondent's answer and the associated weighting that has been defined in the Control table. The key elements of the results table include the following:

Survey Id	Student Id	Goal Id
Control Id	Calculated measure	Other – Please specify
Additional comment		

Learner Profile Rules

One of the objectives of the survey application, specifically dealing with Learning Preferences is the ability to provide an individual who completes the survey with a Learner Profile. A separate table will hold the rules that will be compared to the individual survey results. The key elements of the results table include the following:

Survey Id	Goal Id	No.
From Measure	To Measure	Preference
Comment		

3rd Party Cross Reference / Integration

A separate table will allow key elements of the survey application to be mapped to 3rd party applications. The survey application will collect valuable knowledge about individuals and organisations that could be further enriched by combining the knowledge with other data sources. The key elements of the 3rd party x-ref table include:

Client Id	3rd Party System Code	3rd party X-Ref code

4.4.3 On-line survey entry module design

The design process for this module is outlined using screen mock-ups and story boarding. A hierarchy of JSP files as outlined in figure 4.10 provides the basis for the on-line Survey entry module. Each file contains HTML, embedded JSP and JavaScript.

Note it is not intended to include all of the screen designs in this section. However the design includes a representative number of the most relevant screens.

The online survey application will be launched from an URL attached to an email message or alternatively from an organisation's web site. The student will be presented with a welcome page (figure 4.11) where they can start the survey by entering a valid user id and password. For new users, a register option can be taken which in turn will present the student with a separate page to update the student profile information. See figure 4.12 for the proposed screen design.

At this stage it will be important that the student can record their accessibility preferences and whether assistance is required e.g. authorised scribe, to complete the survey. This screen will contain a number of mandatory entries which when complete, the student can start the survey.

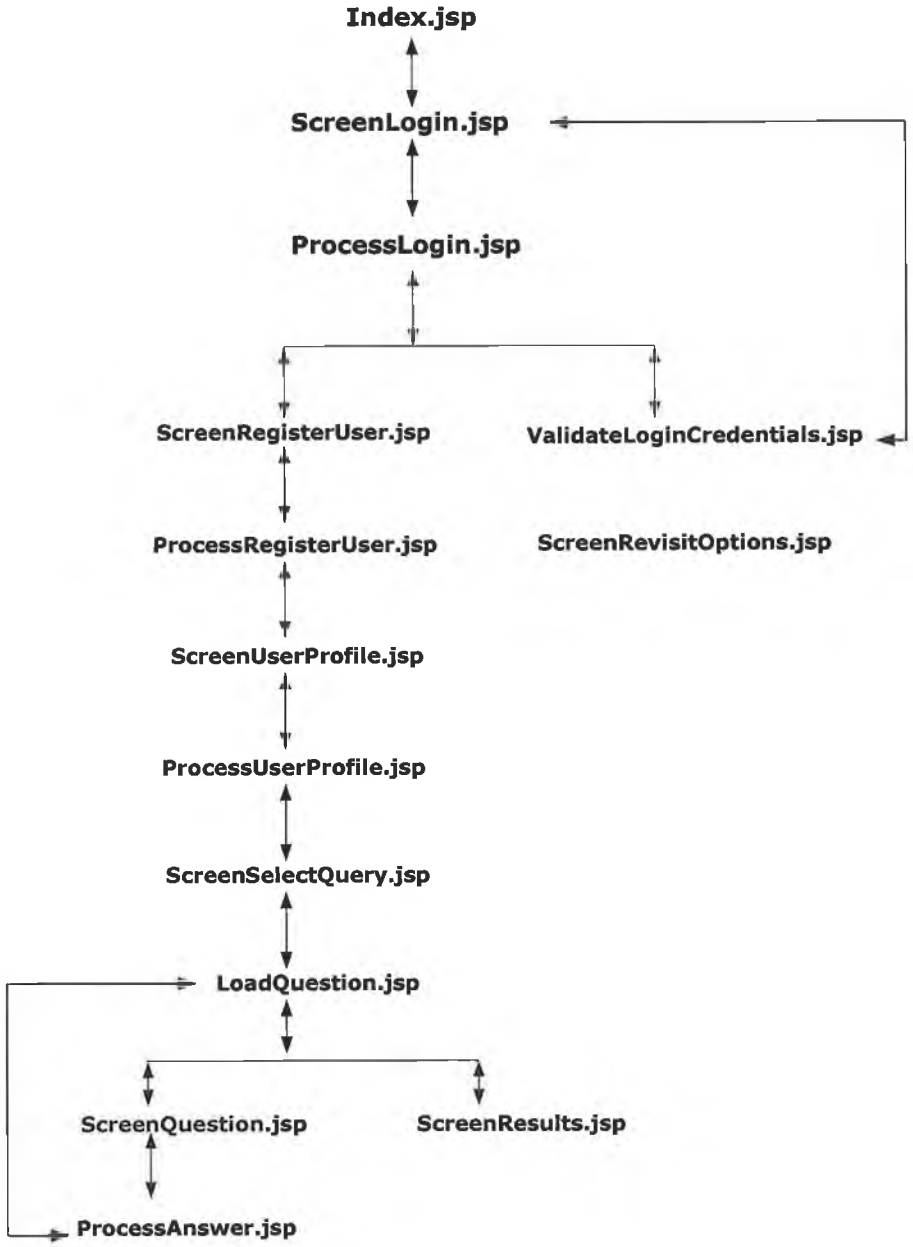


Figure 4.10: Online survey entry web page structure (author, 2005

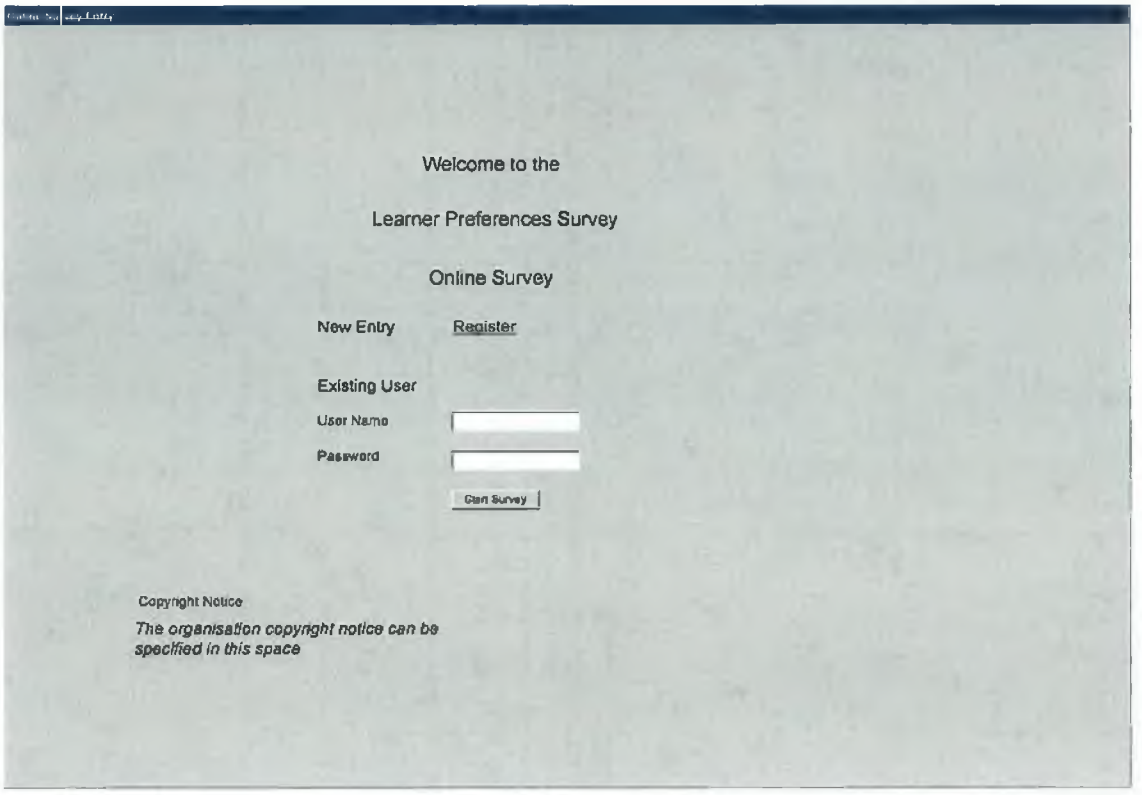


Figure 4.11: Proposed Sign on and security page (author, 2005

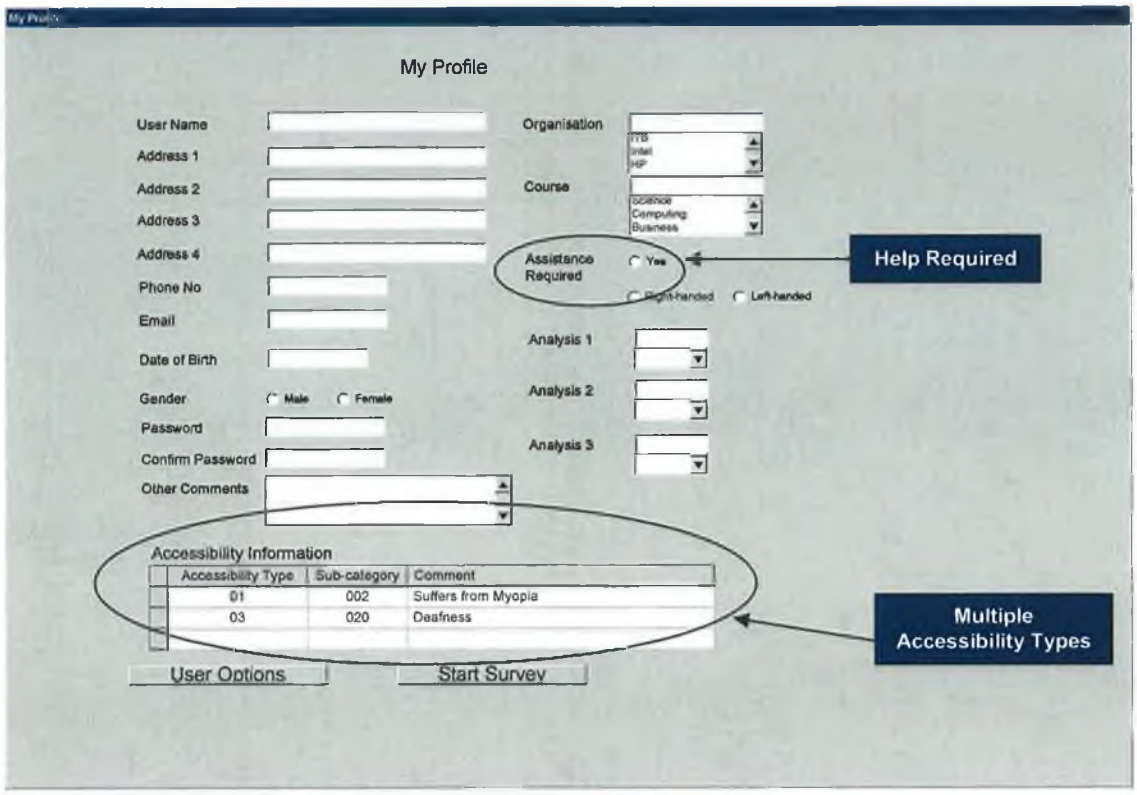


Figure 4.12: Proposed Personal Preferences Screen (author, 2005)

Figure 4.13 contains the proposed design of the survey entry screen. Different formats or question styles will exist for a Control/Question depending on how a Control has been configured. See figure 4.16 relating to Control setup in the Survey Administration section where up to six different question styles can be configured.

Controls are grouped by specific Goal and alternatively a combination of the three Goal Analysis codes. The relevant Goal Id and Goal description will always be displayed at the top of each screen.

When all entry is complete, the student will perform the "Submit survey" option.

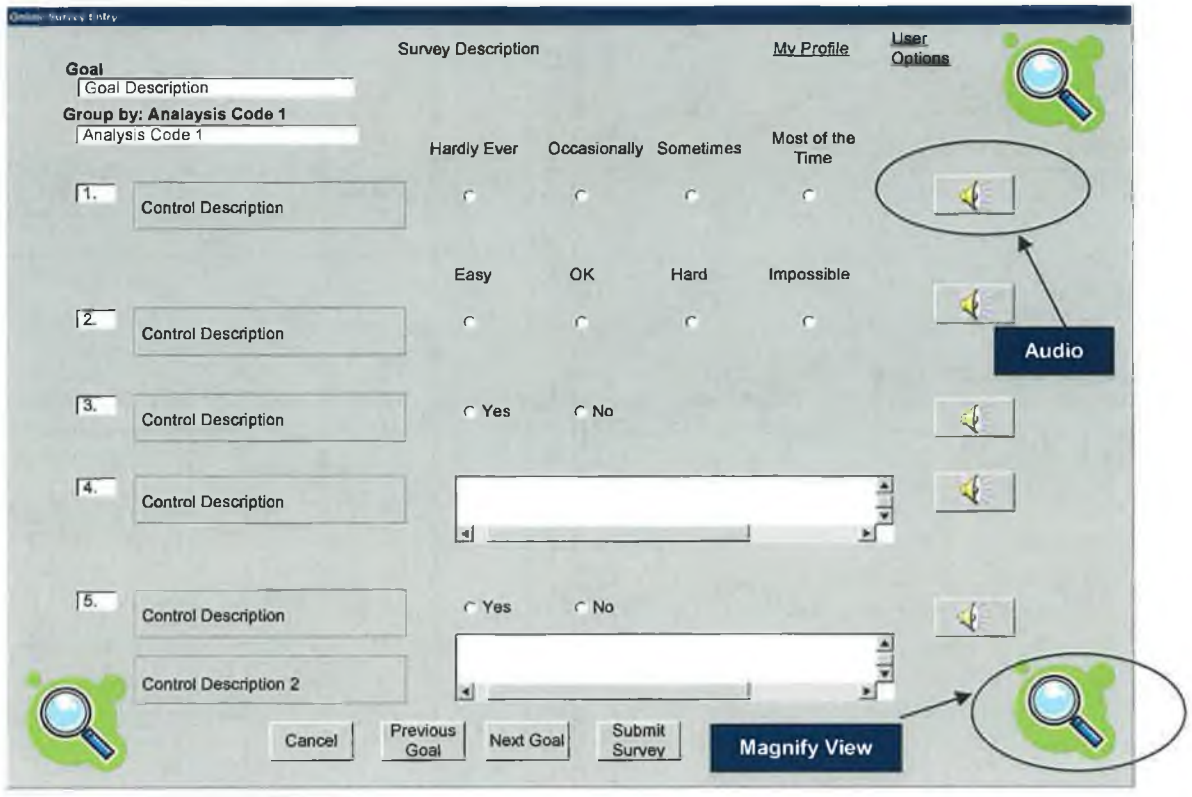


Figure 4.13: Online Survey Entry (author, 2005)

4.4.4 Survey Administration Module

This Survey Administration design process for this module is outlined using screen mock-ups and story boarding. A hierarchy of JSP files as outlined in figure 4.14 provides the basis for the Survey Administration module. Each file contains HTML, embedded JSP and JavaScript. Note that the .jsp pages in the shaded areas represent the pages that will be automatically generated by the DBWings code generator in relation to table maintenance i.e. insert, update and delete. This means that the same number of pages will be repeated for the following .jsp files: ManageCourse.jsp, ManageSurvyes.jsp, DefineCodeTypes,jsp, DefineAnalysisCodes.jsp, DefineGoals.jsp, DefineControls.jsp.

Note it is not intended to include all of the screen designs in this section. However the design includes a representative number of the most relevant screens.

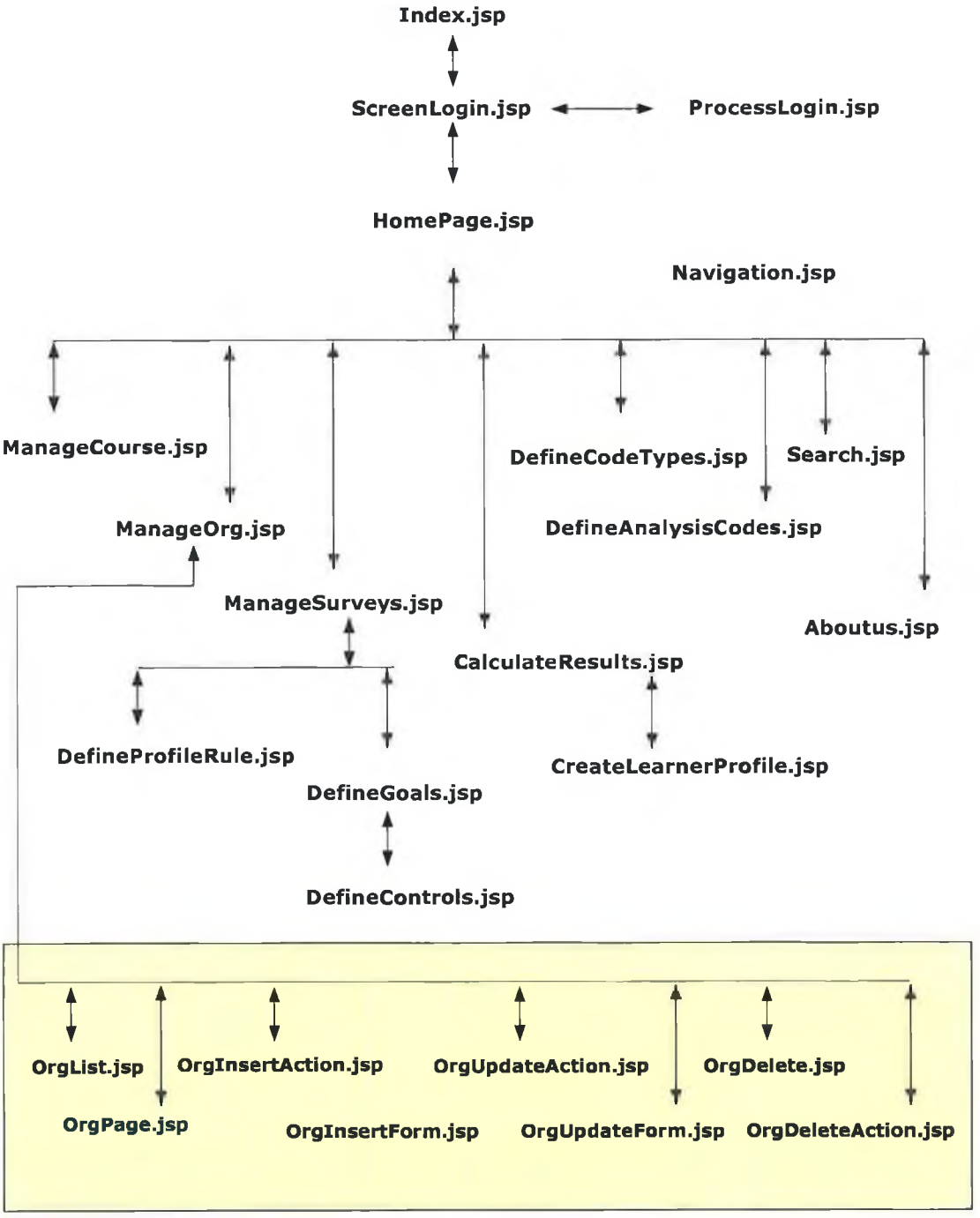


Figure 4.14: Survey Administration web page structure (author, 2005)

Figure 4.15 contains the functions that will be available within the Survey Administration module.

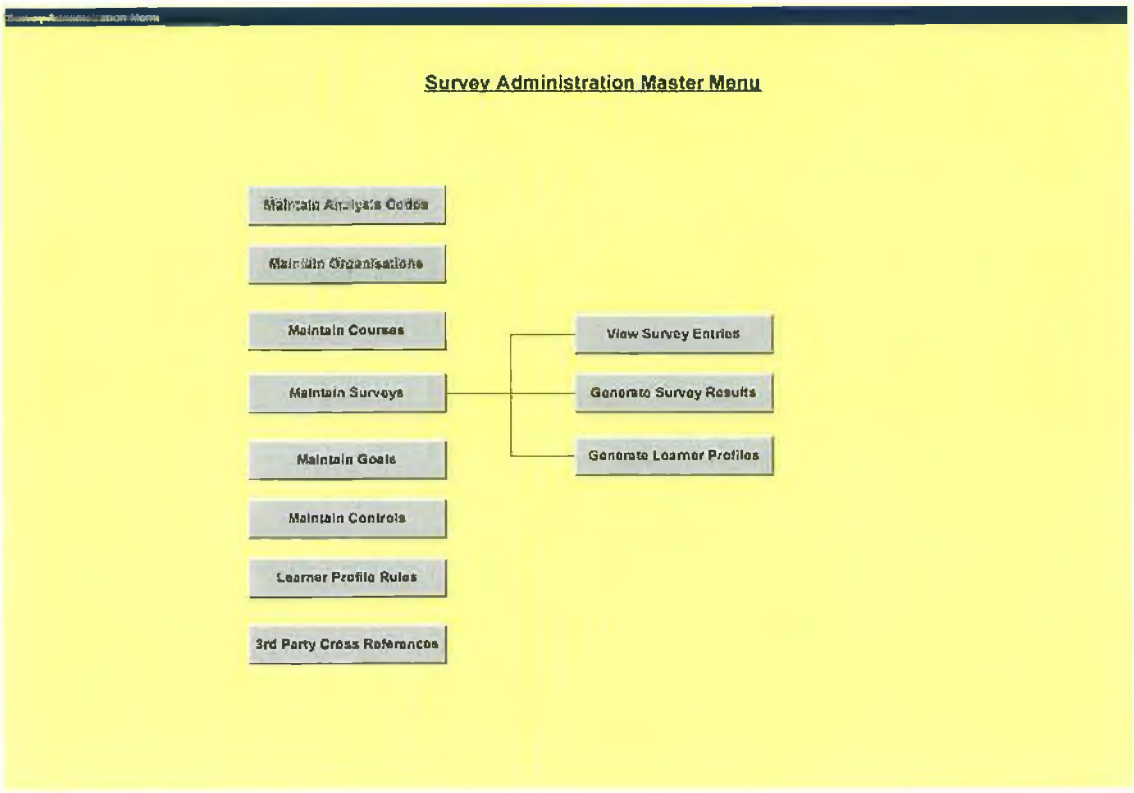


Figure 4.15: Student Administration Master Menu (author, 2005)

The section deals specifically with design aspects that relate to the following areas:

- Control Setup
- 3rd Party Cross References
- Learner Profile Rules
- Generate Learner Profile Report
- Generate Survey Results

Control Setup

The Control setup will be at the heart of the survey application as shown in figure 4.16. Here organisations can configure Controls or Questions within their Goals or Primary groupings. The Control setup design is intended to enable generic survey configurations. For example up to six different question styles are proposed. For multiple choice questions, choice headings can be user defined.

To promote data quality, Controls or Questions can be marked "Answer Required" which ensures that pertinent questions are completed during survey entry.

Unstructured data can also be enabled by allowing an additional comment field that student can qualify their answer during survey entry. As a result, text mining facilities within the Data mining software will be able to generate any relevant trends.

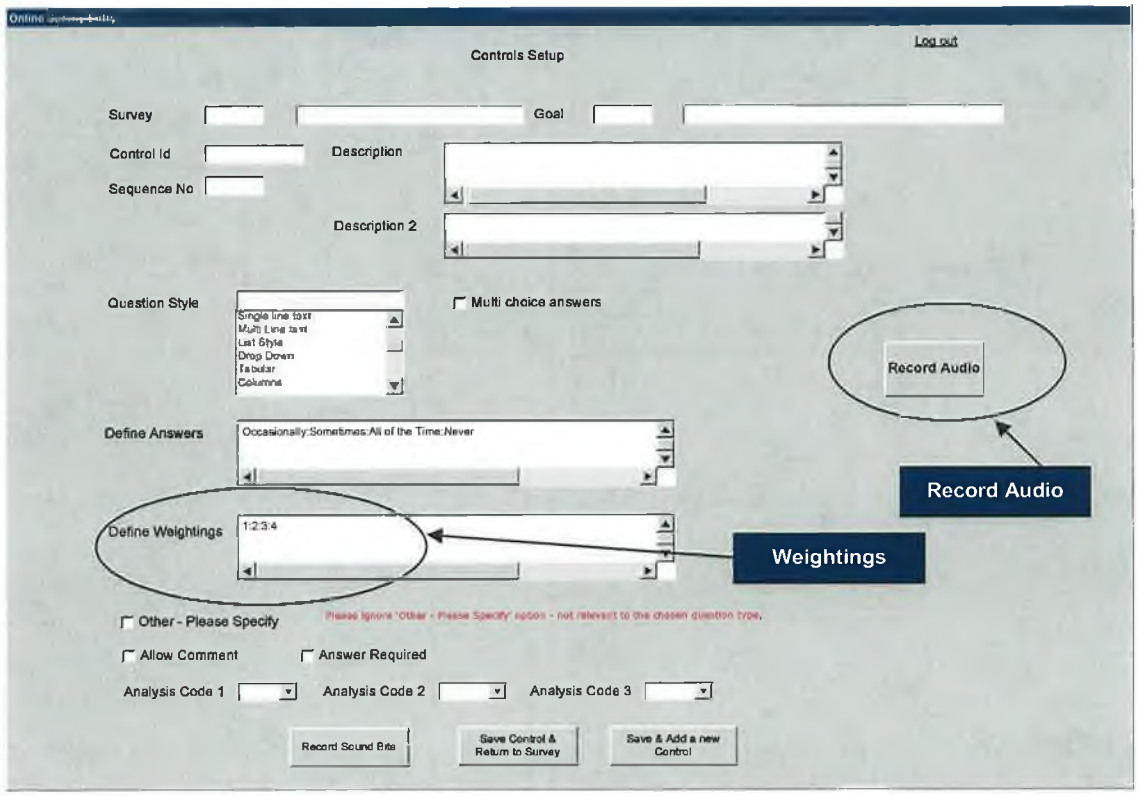


Figure 4.16: Controls Setup (author, 2005)

Third Party Integration

Typically data miners and data warehouse experts are faced with consolidating disparate systems that share common attributes. In the case of Learning Preferences this might include examination results, continuous assessment results, course board student minutes (text mining). It is not unusual that these attributes contain different identifications and as a result difficulties arise in trying to merge common data. During the design of the survey

application system, a number of potential 3rd party integration touch points were identified. These include student and organisation and a sample design is provided in figure 4.17:

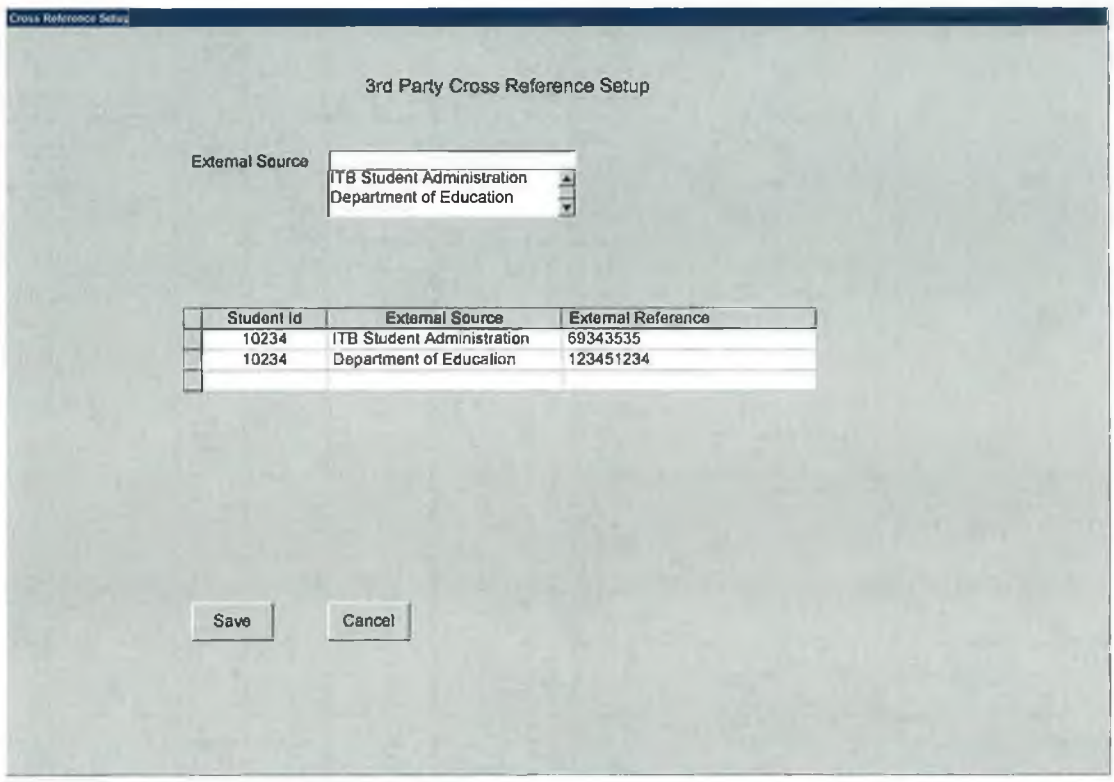


Figure 4.17: Student External Cross References Setup (author, 2005)

In the case of student identification, a separate database table will hold the external cross reference and its associated 3rd party source.

Learner Profile output

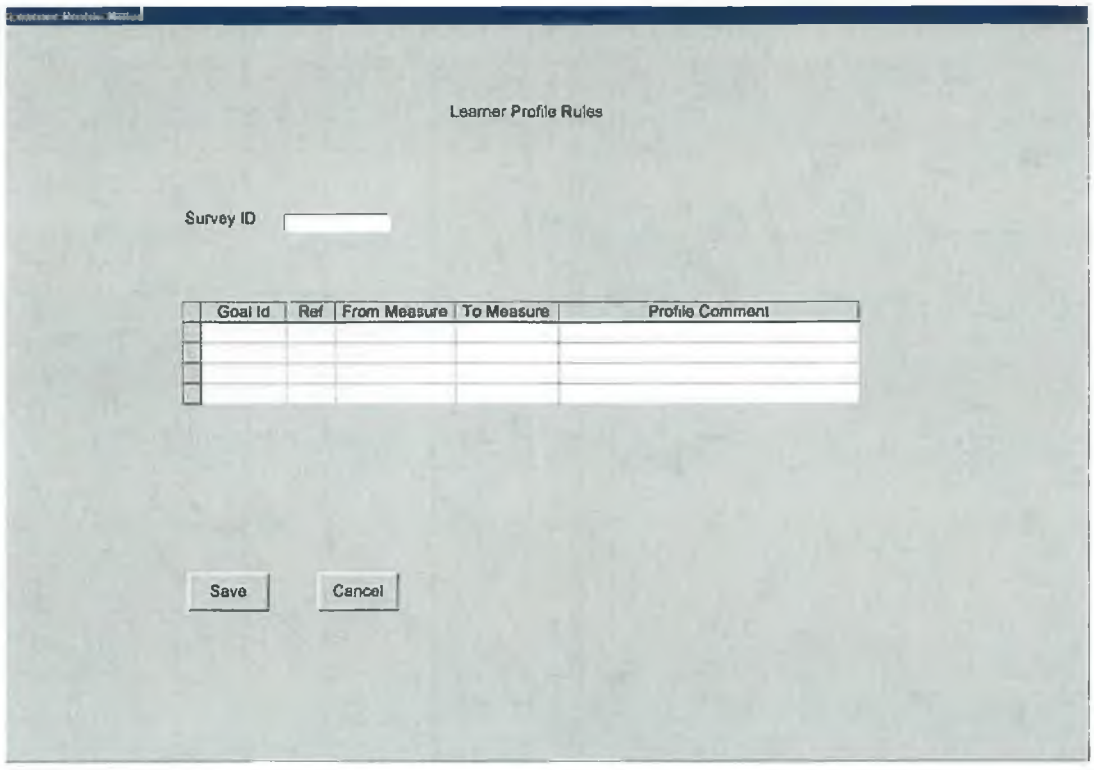
In general, a survey can have two primary areas of interest. Once the survey data has been gathered, it can be analysed and feedback provided to a) the individual and/or b) a company or representative responsible for the collective group of individuals.

The BUA centre is interested in both situations. As previously indicated Data mining will be used to study the total survey data and will attempt to find trends and predictions on a collective basis. A separate process will be developed within the survey application to

produce a Learner Profile for each individual student which will be based on the data input during the online survey.

A number of design features are proposed for the survey application. In figure 4.16, a scoring system for the Learner profile is included in the Control setup. Here a specific weighting will apply to the particular answer or choice of answer that an individual makes to a control/question during the online survey entry.

Learner Profile rules will be maintained on a separate table and the proposed screen design is shown in figure 4.18.



(Screen Design created by the author, 2005)

Figure 4.18: Proposed Learner Profile Rules Screen

The key points to note in relation the design of the Learner Profile report are as follows

- An individual's Learner Profile will be measured by Survey and Goal.

- Each control will contain a measure which in turn is accumulated into a Goal total (see figure 4.16)
- The accumulated Goal and Survey total is compared to their respective Learner Profile Rules
- Each Learner Profile rule will contain a "from and to" measure range which will be used to compare with the individual's accumulated Goal total.
- Each Learner Profile rule will contain an appropriate profile comment that is deemed suitable to the "from and to" measure range
- More than one rule can exist for a combined goal and survey and this will be handled by a separate reference field
- A Learner Profile report will be produced and will contain the following information
    - Personal Details
    - Survey Id and description
    - Goal Id and description for goal defined within a survey
    - Accumulated Goal Measure total
    - Measure Range and Profile Comment

Survey Results

A separate process will be developed to generate survey results. Figure 4.16: Proposed Control Setup screen design, shows where each Control or Question will contain a weighting (in numeric terms). It will be the responsibility of survey administrators e.g. physiologists, will determine the appropriate weighting.

A separate entry will be made to the Survey Results table for each combination of Survey Id, Client Id, Goal Id, and Control Id. An additional field called Measure will contain the relevant weighting value derived from the answer given to the Control or Question during the online survey.

The reason that there is a separate process is to provide physiologists with the flexibility to amend weightings. Therefore there will be the possibility of re-running the process to create new Measure values.

4.4.5 Accessibility Considerations

Accessibility in the Online Survey Entry module

A number of facilities are proposed (figure 4.19) to include a number of properties in the Survey application that will enhance a student's ability to use the application. For example, a user can define display options (font, size and colour options) which may help increase visual contrast. Also, a user can turn on or off sounds that can control the way a student is alerted of particular operations in the software [26].

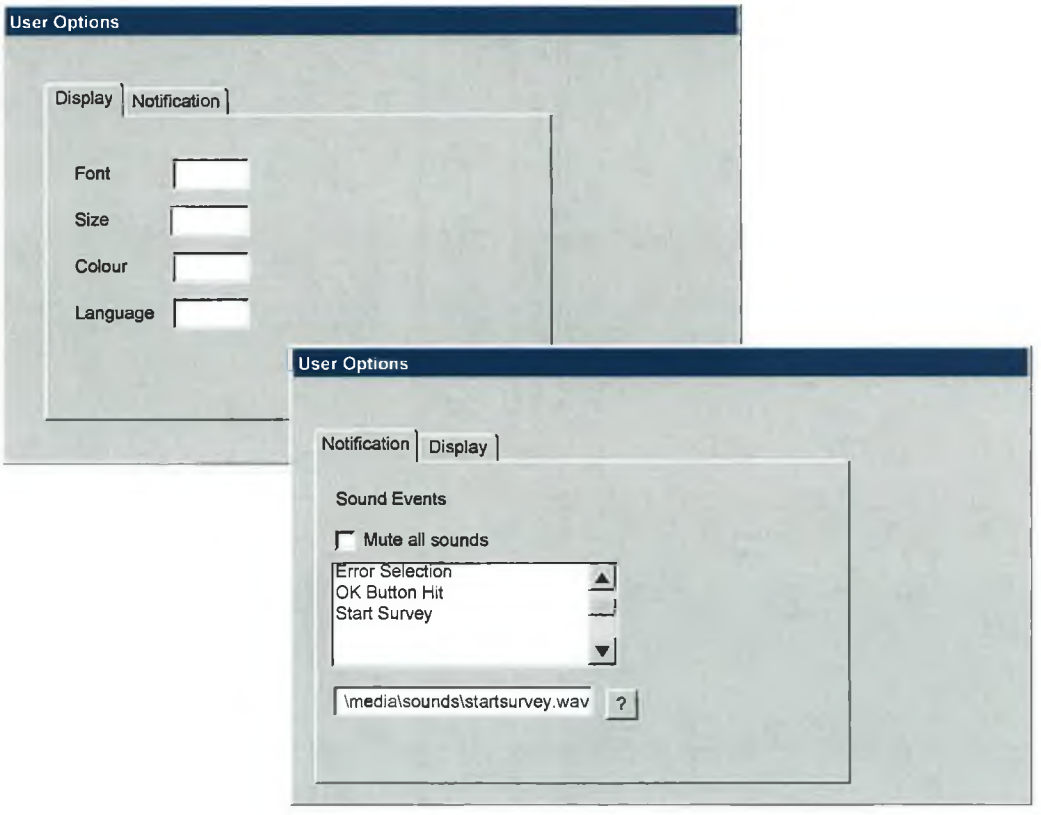


Figure 4.19: User Options (author, 2005)

Analysing Accessibility

During the research stage, it was discovered that an individual can have more than one accessibility type. The current scenario allows a maximum of two i.e. sight and hearing and relies on an individual describing their accessibility in free form comment. In the new survey application, an individual will be able to record multiple accessibility types as illustrated in figure 4.6: Personal Preference screen. In addition, the system will require an individual to indicate a separate accessibility category which will provide another level of accessibility information. For example, an individual might have a sight problem which will be deemed an accessibility type. The particular type of sight problem that the individual suffers from is Myopia which will be recorded in the sub-category field.

An individual must use a valid value that has been previously defined for either field i.e. Accessibility type and Sub-category. This will have significant benefits to data mining in terms of structured data.

4.5 Data Mining

One of the objectives of the Learning Preferences assessment and the online survey application is to satisfy two key principals namely the individual or student who performs the survey and the organisation who sets the survey. From the student's viewpoint, the aim is to provide feedback by way of a learner profile by which the student can learn more about their respective learning strengths and weaknesses. For the organisation, they are very interested in the performance of individuals but equally they are keen to understand about the trends and predictions that the accumulated data from all students presents.

Data mining technology is one of the tools that can help organisations achieve valuable insight using some of its advanced technologies. A number of key questions were raised during the research activity which will be the main focus for the design and implementation of data mining processes alongside the online survey application.

As one of the chief contributors to this research project, the BUA Centre has provided survey data that they conducted and gathered at one of their locations. The data is available in an MS Excel spreadsheet which is an indication of the rather crude method that BUA are currently running and gathering survey data.

This data will be used in two separate data mining scenarios. Each scenario will apply the CRISP-DM methodology to direct the process.

Scenario 1: Data mining will be performed using the BUA spreadsheet.

Scenario 2: Data mining will be performed using the BUA data that has been converted from the BUA spreadsheet into the new survey application database format.

There are number reasons why these two specific scenarios were chosen.

1. Each scenario uses the same data which can be very useful when comparing results.
2. Scenario 1 contains a significant number of unstructured data formats. For example, the responses to hearing and sight have been completed as free form text.
3. Scenario 2 contains data that was previously in an unstructured format that has been converted to a structured format as part of a formal database design.
4. For either scenario, it is hoped that the results produced can illustrate the benefits that data mining can have in relation to survey data, specifically in the area of Learning Preferences

4.5.1 CRISP-DM Methodology

The CRISP-DM methodology, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide data mining efforts.

- As a **methodology**, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.

- As a **process model**, CRISP-DM provides an overview of the data mining life cycle.

The life cycle model consists of six phases as outlined in figure 4.20, with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary. Appendix N contains a detailed breakdown of CRISP-DM phases.

Figure 4.20: CRISP-DM Life Cycle (SPSS, 2005

CRISP-DM allows for the fact that with most projects it is often necessary to return to earlier phases in the project. In fact CRISP-DM actively encourages this movement. This research project is looking at data mining in terms of where it can add further value to survey applications and indeed the benefits of considering data mining during the design of the survey application.

The CRISP-DM model is flexible and can be customised easily. For example it is not required to have a specific modeling goal which means the focus can be on data exploration and visualisation to uncover suspicious patterns in population data. CRISP-DM allows the creation of a data mining model that fits the particular needs. For this research project, the

Data Understanding and in particular the Data Preparation phases of the project will be the most important, as it is expected that these phases will provide valuable information in relation to key elements of the research objectives. CRISP-DM allows for the project to be approached in such a way that it will follow the methodology but will also be structured in such a way as to focus on the most relevant phases.

4.5.2 Clementine Data mining software

The Clementine software contains two key elements that play a vital part for all data mining processes. These include the Stream and Node and an example is shown in figure 4.21 and a brief explanation is provided below:

What is a Stream?

Everything that is performed to the data is done within a stream in Clementine [26].

- The stream is a blank canvass for the user to work on
- A stream can be used to achieve different goals or one stream can be used to perform a number of tasks
- Tasks in a stream are executed in a specific order, a stream only travels in one direction

What is a Node?

A node is task that can be added to a Clementine stream [26].

- A node can take in data, output data or allow data to pass through it
- There are six category of nodes
  - Source Nodes – used data to bring data into Clementine
  - Record Operations – used to manipulate records
  - Field Operations – used to manipulate the number fields in the dataset
  - Graph Nodes – obtain a graphical representation of the data
  - Modeling Nodes – tools to construct a model
  - Output Nodes – used to analyse, report and export data

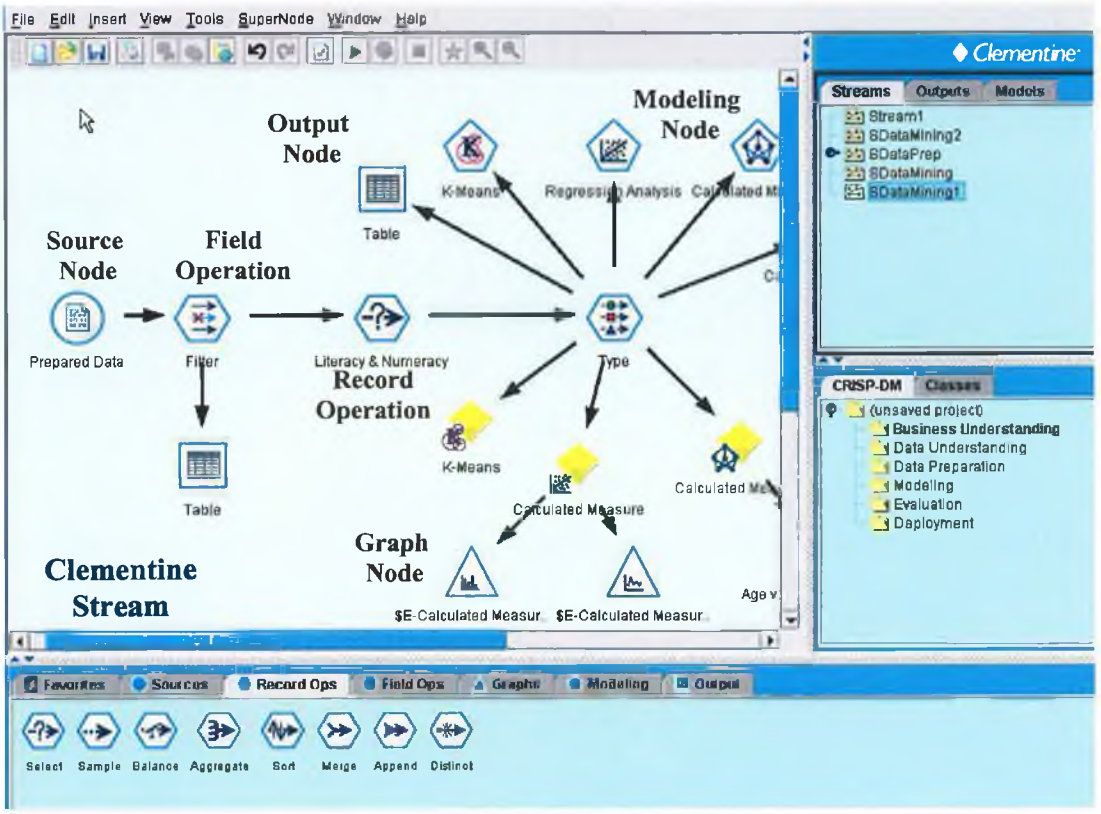


Figure 4.21: Example of a Clementine Stream and related Nodes (SPSS, 2005)

4.5.3 Scenario 1: Data mining using the BUA spreadsheet.

Business Understanding

Even before working in Clementine, it is very important to take the time to explore what the organisation expects to gain from data mining. This is typically performed as part of the Business Understanding phase of CRISP-DM.

When clearly defined goals are established, an assessment is typically made in relation to the quality of the data that is available for analysis. For this scenario, the focus was around the BUA spreadsheet. BUA had no real expectation in relation to the data that it provided due to the weakness of their existing systems. However they were interested in terms of any unknown knowledge that data mining might throw up.

Once clear goals have been identified, it is important to translate the organisational goals into data mining goals. Because of the unknown nature of the data and lack of clear goals, a Clustering or Unsupervised Learning strategy was identified as a primary focus.

Data Understanding

This is the stage where the data miner gets familiar with the data they are mining. The major areas of note that the data miner must understand in this phase include the following:

- Explore and define what each field in the dataset presents
- Be familiar with the attributes of the values within the fields
- Check whether alpha or numeric values are stored
- Check visually for interesting or unusual values e.g. missing values

Dataset Description

The BUA dataset contains a MS Excel spreadsheet which represents a summary of the Learning Preferences survey conducted by the BUA Centre.

- The population dataset contains 95 records which comprises one record for each student who completed the learning preferences survey. Note that each record contains five measures which represent unique learning preferences. These measures have been determined by BUA using a manual process.
- Each unique measure represents a calculation based on the chosen answer from each of the questions posed to the student during the survey.
- In order that the BUA data can be used by the Clementine data mining software it is necessary to convert the spreadsheet into CSV (comma separator variable) format. A CSV file is also known as a flat file which is one in which table data is gathered in lines of ASCII text with the value from each table cell separated by a comma and each row represented with a new line.

Dataset Specifics

The dataset specifics that are typically discovered during this phase include the following:

- Verify the data quality
- Which fields are essential
- Which fields offer nothing to the data mining project

- Which fields complement each other
- Is there enough relevant data to continue
- Are there missing values
- Will they cause a major problem to the mining

One of the first steps within this phase of CRISP DM was to get a better understanding of the data being modeled. Figure 4.22 outlines how a Clementine stream has been defined to explore the data being modeled. A 'Quality' output operation node is used to highlight data quality in terms of missing values or blanks and the results are shown in the example in Figure 4.23

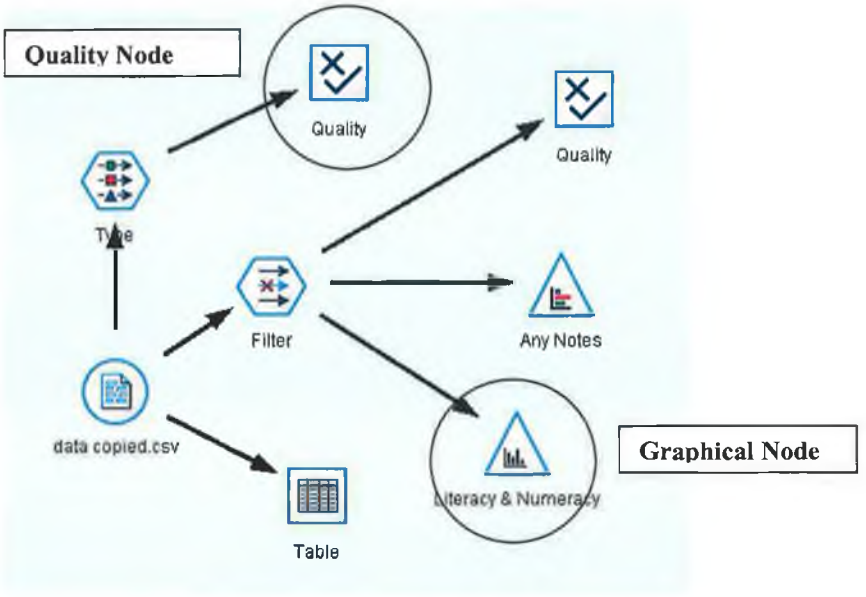


Figure 4.22: Clementine Output and Graphical Nodes (SPSS, 2005)

The output from the quality node is presented by the table in figure 4.23 which is summarised as follows:

- Some fields contain free form text. For example, the 'Sight' and 'Hearing' fields contain free form text which can be inconsistent and invariably misspellings will occur.
- This operation should be able to isolate fields that contain missing or blank values.
- Some fields will contain non-numeric values where numeric ones are expected.

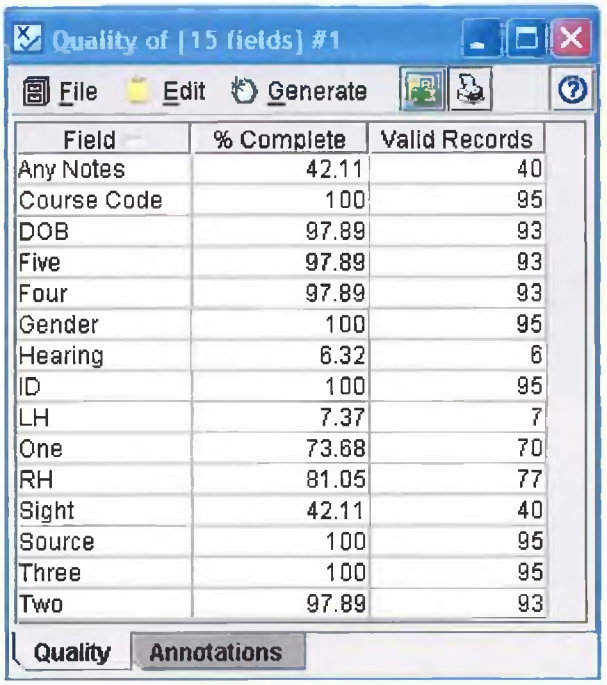


Figure 4.23: Output from the Quality Node (SPSS, 2005)

- Clementine provides other features that facilitate the visual representation of the data. There are several graphical output nodes which can quickly illustrate data. Figure 4.24 demonstrates a histogram of how student performed under one of the learning preferences measures. The example in this case is Literacy & Numeracy.

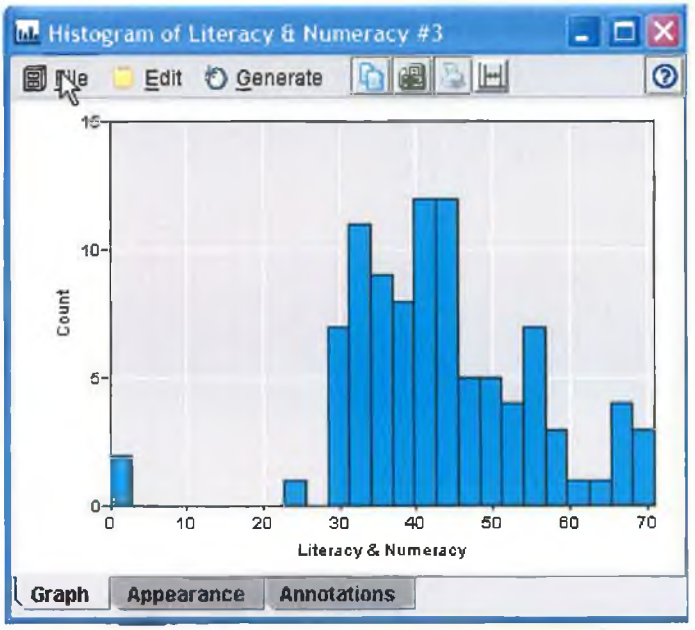


Figure 4.24: Output from the Graphical Node (SPSS, 2005)

Data Preparation

Data preparation is the process where the data available is prepared and structured in such a manner as to achieve the greatest results when it is data mined. There are several techniques used during this phase, the most common being:

- Data cleansing
- Derivation of new fields

Using SuperNodes

Clementine contains a nice feature to remove some of the nodes from the primary stream into one or more related streams called "☆ SuperNodes". This feature is very useful when a large number of nodes are used and grouping related nodes together is preferred. It also means that visually, the stream is easier to read and understand. In this scenario a "*SuperNode*" is proposed to define and manage the data preparation of specific fields using the 'Derive' field operation node.

In Clementine, it is possible to click on the SuperNode, shown in figure 4.25 and the system will automatically display a second stream that contains related nodes. An example is shown in figure 4.26.

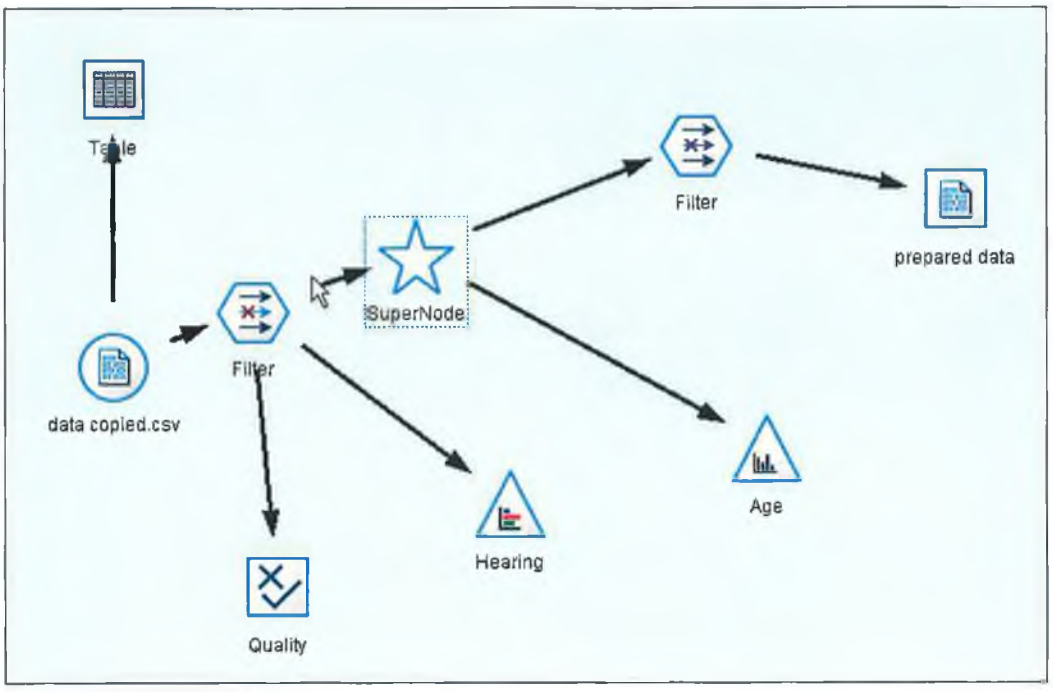


Figure 4.25: Example of a SuperNode (SPSS, 2005)

The SuperNode contains a number of Field Operation nodes that repairs data quality issues which are inherited from the source data e.g. data warehouse, spreadsheet, etc. A 'Derive' operation node is used to transform the following fields which in turn will be used by further data mining operations such as the modeling algorithms. The fields that have been included are:

a) Left or Right Handed

b) Sight

c) Hearing

d) Age

In the case of a), b) and c) the point is clearly illustrated that data accuracy problems exist and highlight the importance of data preparation for data mining success. Look at figure 4.27 which shows how Clementine converts the free form data that was used to record the type of sight the student had. Similar issues arise for Left or Right handed and Sight fields. In the ideal situation, predefined values would be the best approach which is the aim of the new Survey Application.

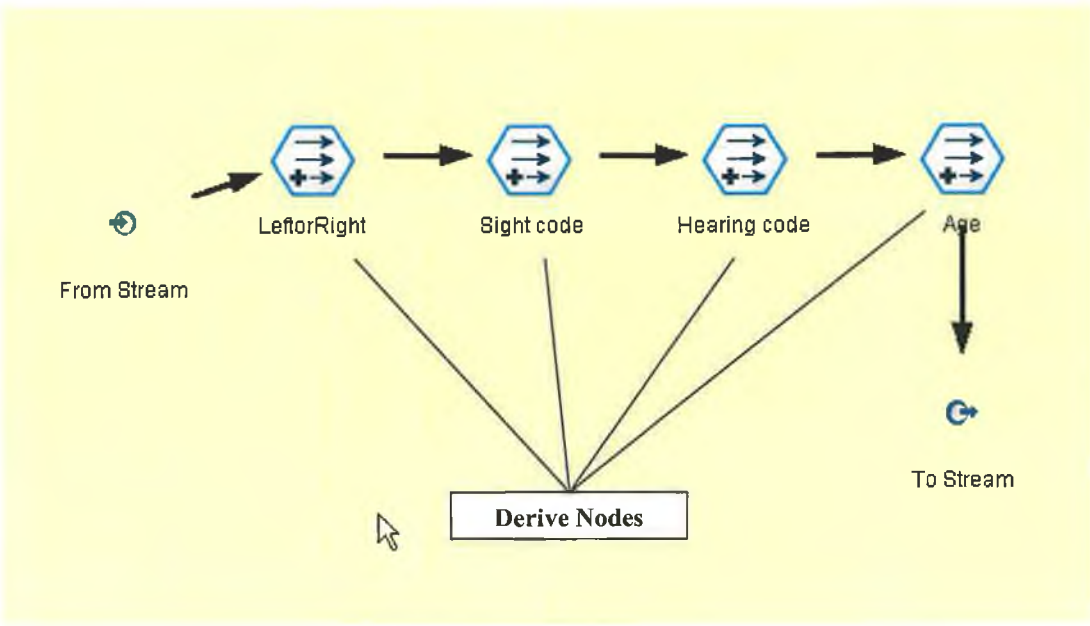


Figure 4.26: Example of a related SuperNode stream (SPSS, 2005)

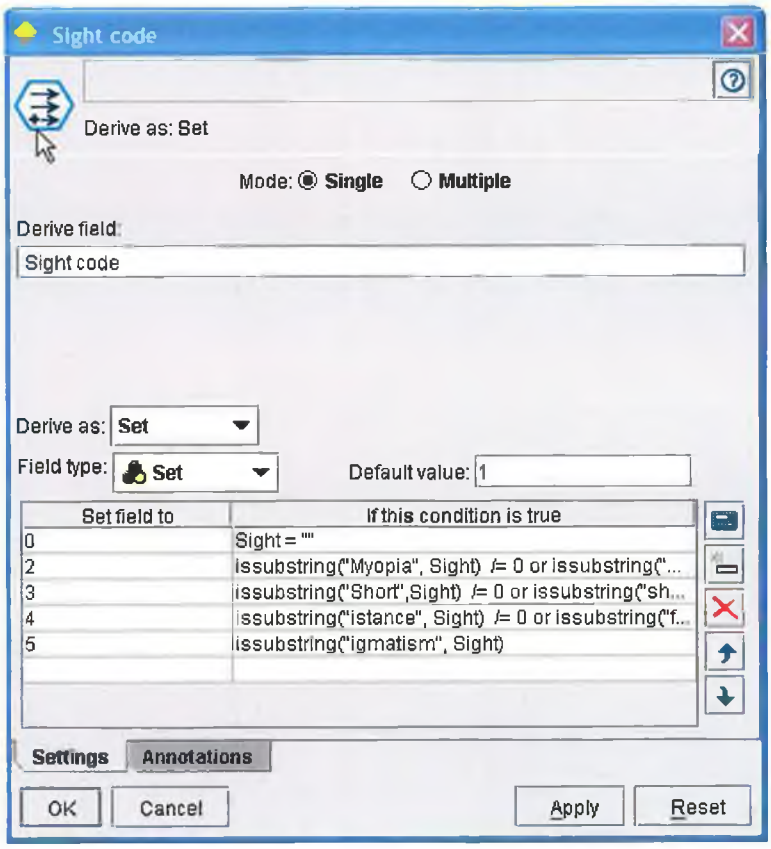


Figure 4.27: Derive Node operation (SPSS, 2005)

Modeling

During the Business Understanding phase, a Clustering or Unsupervised Learning strategy was identified as a primary focus. Three modeling algorithms (K-Means, TwoStep, and Kohonen) which handle clustering were selected and shown in the Clementine stream in figure 4.28. The objective was to establish whether clustering could determine any trends or patterns in the data that was prepared during the previous stages.

In figure 4.29, the K-Means model identified 5 clusters by establishing a grouping of following attributes Age, Attention & Concentration, Co-ordination, Social & Communication, and Education History. The other algorithms generate different cluster outcomes.

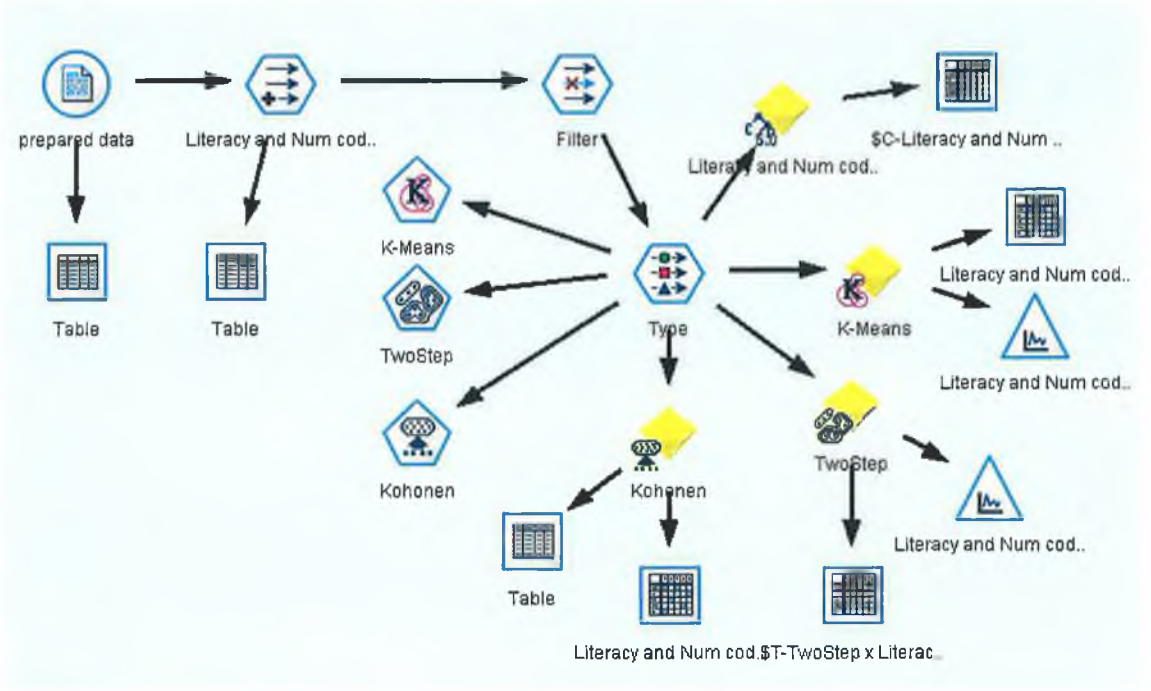


Figure 4.28: Example of cluster models applied to Scenario 1 (SPSS, 2005)

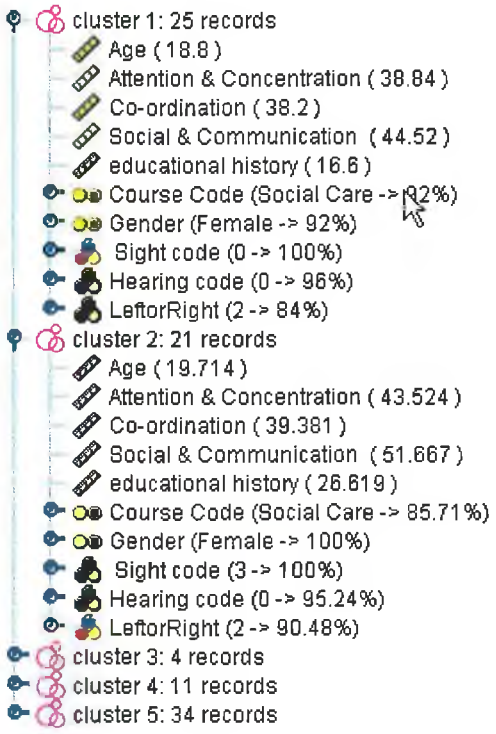


Figure 4.29: K-Means clusters (SPSS, 2005)

4.5.4 Scenario 2: Data mining using the new survey application database structure.

Data mining using the BUA data that has been converted from the BUA spreadsheet into the new survey application database structure.

The objective of this scenario is as follows

- Apply the CRISP-DM methodology to focus on data exploration and visualisation to uncover suspicious patterns in the population dataset
- To achieve the greater success under the terms of the research objective when compared to scenario 1
- Illustrate the potential for further data mining scenarios

A significant step is required before performing data mining for this scenario. This involved the conversion of the BUA Spreadsheet into the new Survey Application database

design as per section 4.4.2. Other points to note in relation to the conversion include the following:

- The conversion was dependent on the completion of the MySQL database design for the new survey application.
- The conversion of the BUA spreadsheet into the survey MySQL database using MySQL freeware tools [27]
- The converted survey did not contain a full set of survey data because the survey was not performed using the online survey system
- As a result of the conversion, it was generally felt that there were significant improvements in the scope and quality of the data

Business Understanding

Even before working in Clementine, it is very important to take the time to explore what the organisation expects to gain from data mining. This is typically performed as part of the Business Understanding phase of CRISP-DM.

When clearly defined goals are established, an assessment is typically made in relation to the quality of the data that is available for analysis. For this scenario, the focus was around the BUA spreadsheet that had been converted to the new survey application database. BUA had no real expectation in relation to the data that it provided due to the weakness of their existing systems. However they were interested in terms of any unknown knowledge that data mining might throw up.

Once clear goals have been identified, it is important to translate the organisational goals into data mining goals. Because of the unknown nature of the data and lack of clear goals, a Clustering or Unsupervised Learning strategy was identified as a primary focus

The availability of the survey application database offers significant opportunities to an organisation and the data miner. By virtue of examining the data model in figure 4.30, it is possible to visualise the layers of knowledge that can be utilised. The Survey Responses, Learner Profile and Results information is at the heart of the data model. This layer is

surrounded by a set of master data that can be tightly integrated with a facility to link to 3rd party data sources.

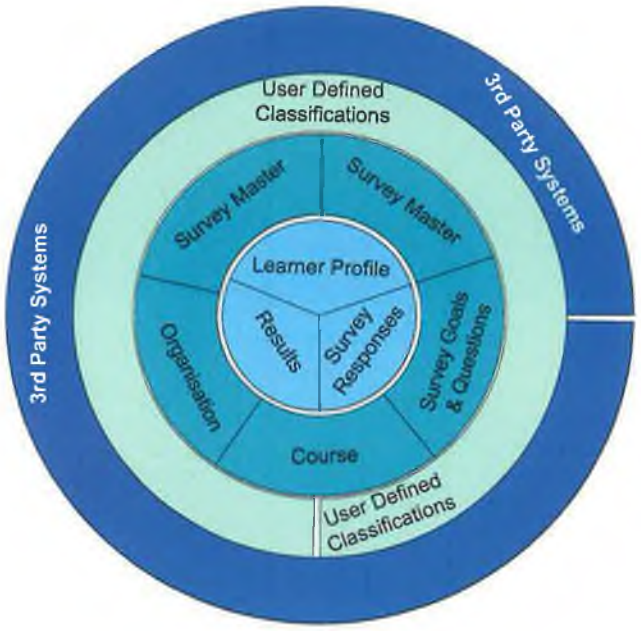


Figure 4.30: Survey Application data model (author, 2005)

Data Understanding

Dataset Description

The BUA dataset consists of one MS Excel spreadsheet that was converted to a MySQL database represents a summary of the Learning Preferences survey conducted by the BUA Centre. Other information includes the following:

The population dataset

The population dataset contains 475 records which consist of 5 sets of measures for 95 students. In the first test the dataset contained 95 records and the each record included the five measures.

It was initially felt that some data preparation between the survey database and Clementine would be required using SQL. While this is perfectly okay to do, it was subsequently discovered that it is not necessary as Clementine provided in-built facilities to select and

merge external database tables. The initial design of the Survey application contained a facility that would generate its own SQL to extract user defined datasets.

This scenario uses database tables as the data source for data mining. The new survey application contains several options in terms of data mining opportunities. For the purpose of this scenario, the following tables will be used in conjunction with a Clementine merge node:

- Student Preferences
- Student Accessibility
- Results

Initially when the merge node, as shown in figure 4.31, was used to join the above database tables, Clementine did display a key field by which tables could be merged. Subsequently it was discovered that that the key fields that need to be joined between tables must contain the same field name. This can be achieved by using the filter tab on the database node whereby the column name can be overridden.

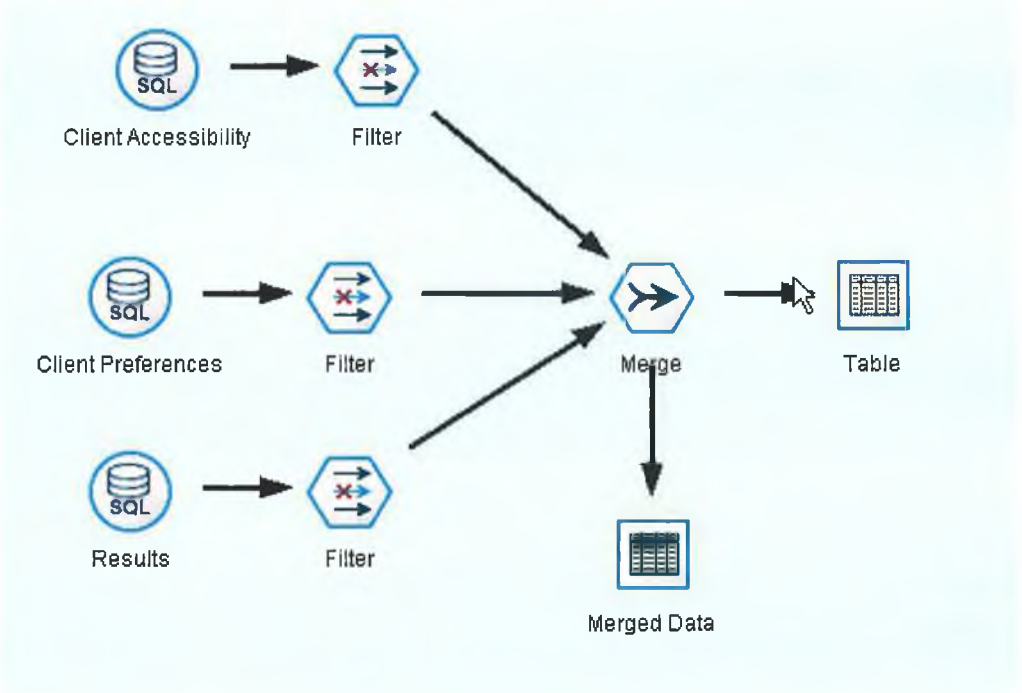


Figure 4.31: Merge MySQL database tables (SPSS, 2005)

The client preferences table contains a record for each student who completes the survey. The columns included and selected as part of the Clementine filter node are:

Client Id	Organisation Id	Date of Birth
Gender	Course Id	Left or Right Handed
Assistance Required	Notes	

The client accessibility table contains a list of all of the accessibilities that relate to a student. Importantly, the new design allows an unlimited number of entries to be recorded. The columns included and selected as part of the Clementine filter node are:

Accessibility Category	Accessibility Sub-Category	Comment

The Result table contains the measure determined by the system in relation to the student's answer to the survey question. A separate program in the survey system generates the measure based on a set of rules. The columns included and selected as part of the Clementine filter node are:

Survey Id	Client Id	Goal Id
Control Id	Result Date	Calculated Measure
Specific Comment	Other Comment	

The output from the merge process produced a merged table somewhat similar to the data source used in the first test involving the use of the BUA spreadsheet. However there are some significant differences in terms of how the data is organised and presented.

In the spreadsheet all of the client data was maintained in one row. In this scenario, a unique row exists for each type of measure. Note that there are five measures in the learning preferences survey which include Education History, Literacy & Numeracy, Social & Communication, Attention & Concentration and Co-ordination. The merged table contains the following columns:

Survey Id	Client Id	Goal Id
Control Id	Result Date	Calculated Measure
Specific Comment	Other Comment	Organisation Id
Date of Birth	Gender	Course Id
Left or Right Handed	Assistance Required	Notes
Accessibility Category	Accessibility Sub-Category	Comment

Data Preparation

A data preparation stream, as shown in figure 4.32, was created in Clementine to clean the data. Some of the nodes have been encircled in order to emphasise a further explanation. Firstly "*Merged data.dat*" contains the merged table prepared from a previous stream shown in figure 4.31. The "*SuperNode*" can be expanded into a related stream, as shown in figure 4.33, to perform the data cleaning for this dataset. The "*Prepared Data*" node is the Flat File Node which allows data to be written to a delimited text file. This is useful for exporting data that can be read by other analysis or streams.

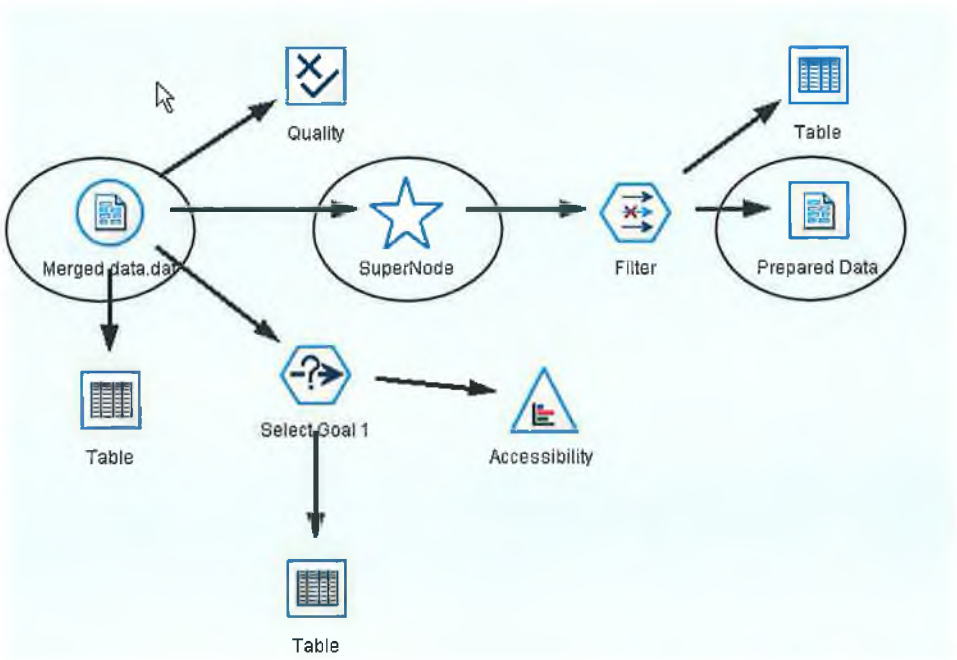


Figure 4.32: Data Preparation Stream (SPSS, 2005)

The Supernode in figure 4.33 contains several Derive Nodes which are used to modify data values and derive new fields from existing data. Generally, Data mining models work better with numeric values and therefore derive nodes are used to convert non-numeric values where possible. This is illustrated in figure 4.34 to figure 4.38.

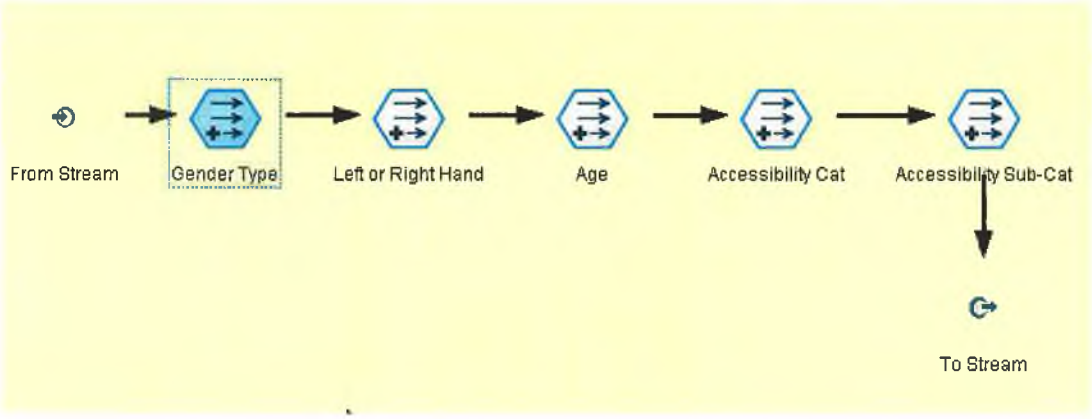


Figure 4.33: Data Preparation SuperNode stream (SPSS 2005)

Set field to	If this condition is true
0	Gender = 'Female'
1	Gender = 'Male'

Figure 4.34: Derive Gender values (SPSS, 2005)

Set field to	If this condition is true
2	'Left or Right Handed' = 'Left'
3	'Left or Right Handed' = 'Right'
0	'Left or Right Handed' = ""

Figure 4.35: Derive Left or Right Handed (SPSS, 2005)

If:

@NULL('Date Of Birth') or 'Date Of Birth' = ""

Then:

22

Else:

datetime_year(datetime_now) - datetime_year('Date Of Birth')

Figure 4.36: Derive Age values (SPSS, 2005)

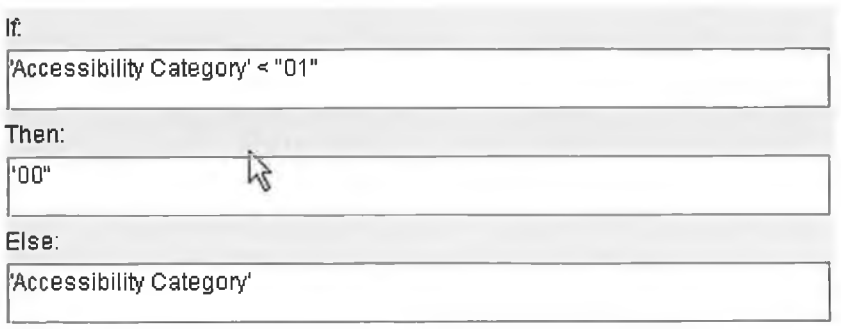


Figure 4.37: Derive Accessibility Category values (SPSS, 2005)

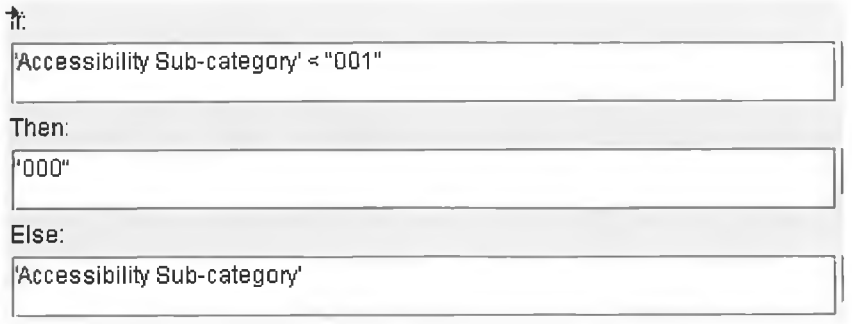


Figure 4.38: Derive Accessibility Sub-category values (SPSS, 2005)

Modeling

During the Business Understanding phase, a Clustering or Unsupervised Learning strategy was identified as a primary focus. Three modeling algorithms (K-Means, TwoStep, and Kohonen) which handle clustering were selected and shown in the Clementine stream in figure 4.39. The objective was to establish whether clustering could determine any trends or patterns in the data that was prepared during the previous stages.

Of the three algorithms used, K-Means produced the most visible results. In figure 4.40, the K-Means model identified 5 clusters by establishing a grouping of following attributes Goal, Age, Accessibility Category and Accessibility Sub-Category.

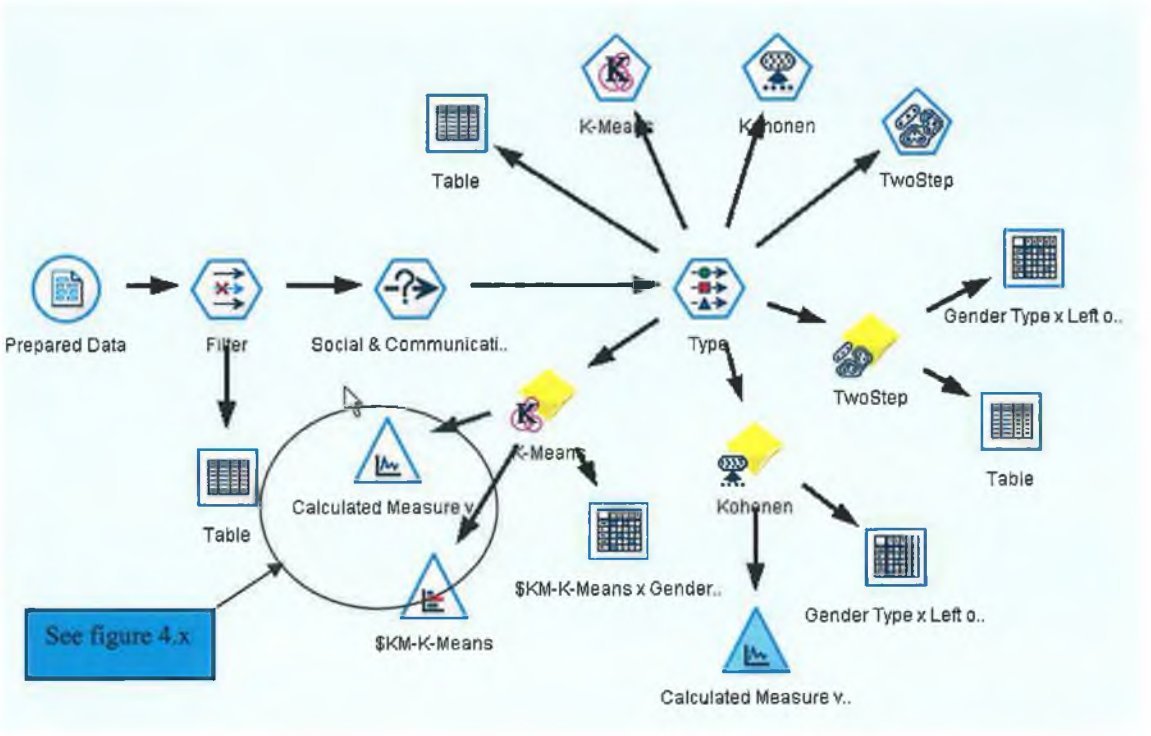


Figure 4.39: Example of cluster models applied to Scenario 2 (SPSS, 2005)

The scatterplot in figure 4.41 represents the clusters in a graphical format. The model has identified a natural grouping of age and accessibility category as a potential trend.

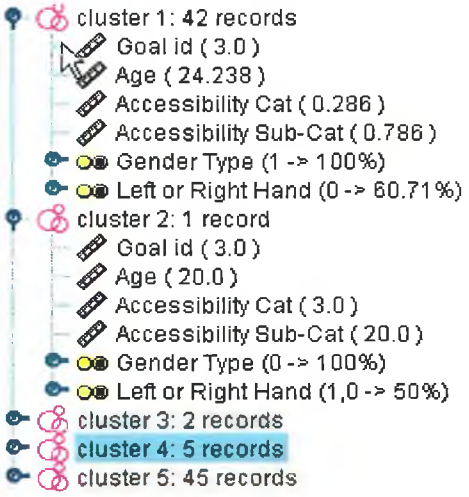


Figure 4.40: Example of clusters applied to Scenario 2 (SPSS, 2005)

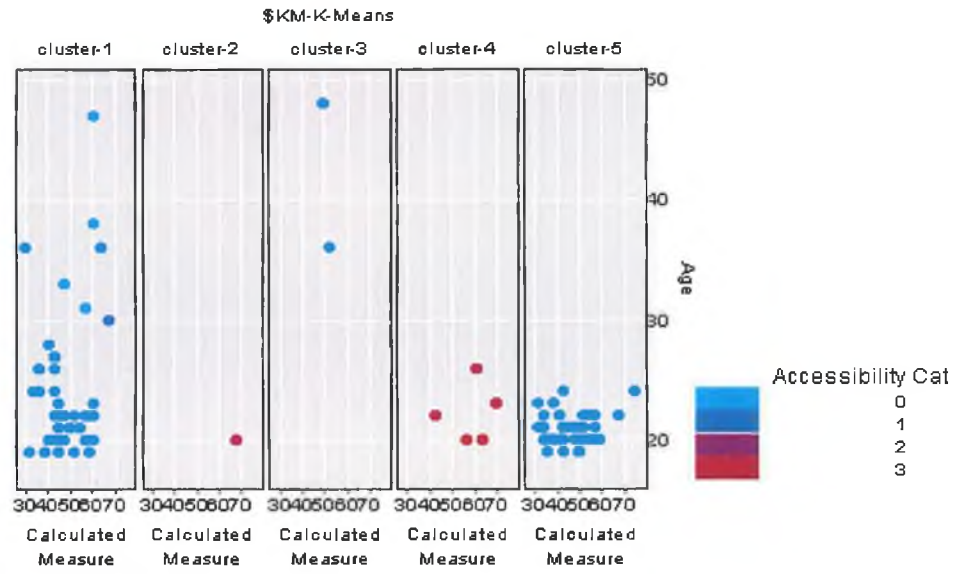


Figure 4.41: Scatterplot diagram created by K-Means model (SPSS, 2005)

4.6 Chapter Conclusion

The solution design incorporates the research aims; drawing together the areas of survey research, accessibility and data mining into a technology framework. A key design factor is the ability to integrate a number of technologies. The database design is crucial to ensuring that a flexible survey application is provided but equally the richness of its attributes will optimise the data mining potential. The next chapter aims to evaluate the results of meeting the research aims through design and development.

5. Evaluation & Testing

5.1 Chapter Summary

This chapter outlines the evaluation and testing strategy that was performed in conjunction with this research project. The strategy is broken into four main areas:

1. Testing approach in relation to the software development of the online survey application
2. Testing approach performed in relation to data mining.
3. Evaluation of research project objectives
4. Other evaluations

Each of the areas required a different approach in terms of the methods that were applied. For example, software development follows industry standards in terms of testing. For data mining, the concept of a train and test dataset was used, where the data miner uses the train dataset to become familiar with the data and the test dataset is used to perform the test in which the results and outcomes are reported upon.

The evaluation of the research objectives, focused on specific areas that have been elaborated upon during the research and design chapters. In some cases the results were based on observations that are contained in the proposed design as outlined in chapter 4.

5.2 Testing Approach for Software Development

The Software testing approach that was applied to the development of the web based survey administration module and online survey entry module follows traditional testing procedures that can be broken into the following areas:

- **Unit testing** involves the testing of individual pieces of development.
- **Integration testing** involves testing one or more pieces of development to ensure it integrates correctly as a collective unit.
- **System testing** involves testing all areas of the system i.e. end to end test incorporating the Survey Administration and the online Survey entry system.

5.3 Testing Approach for Data Mining

Before building the model, it is important to consider in advance how the model's results will be tested. Typically, there are two parts to generating a comprehensive test design:

- Describing the criteria for the "goodness" of a model
- Defining the data on which these criteria will be tested

A model's goodness can be measured in several ways. For supervised models, which incorporate algorithms such as C5.0, GRI, and C&RT, measurements of goodness typically estimate the error rate of a particular model. For unsupervised models, with algorithms such as K-Means, Kohonen cluster nets, measurements may include criteria such as ease of interpretation, deployment, or required processing time [26].

Model building is an iterative process. This means that tests on the results of several models will be performed before deciding on the ones to use and deploy.

5.4 Evaluation Approach for Research Questions

In section 1.4 of this document, the underlying objectives of the research project are identified. In order to measure the success of the research and subsequent implementation of the framework design, it is important that these research objectives are evaluated. For evaluation purposes, the objectives have been further broken down.

Objective 1: How can the design of an accessible system address the problems of Survey Research data?

What are the main accessibility technology issues that enable inclusive data collection?

WAI-TIES have designed a template document presenting a recommended format for communicating results of the evaluation of Web site accessibility according to Web Content Accessibility Guidelines (WCAG) 1.0. A consistent and comprehensive evaluation report format can help ensure effective evaluations as well as accurate

comparisons of accessibility levels over time and between different versions of the Web site. (Appendix M).

Test Case	**Test with BOBBY** **Bobby is a WCAG Assessment Tool. This tool mechanically analyses individual HTML pages, or groups of such pages, for conformance with the WCAG guidelines.**
Reason for Test	To evaluate conformity with WCAG 1.0 guidelines
Testers	Margaret Kinsella
Scope Limitations	The availability and license of BOBBY technology
Result	A basic test was performed.
Comment	Further testing will be required

Test Case	**Trial with JAWS Speech User Agent**
Testers	Centre for Inclusive Technology, NCBI
Reason for Test	To evaluate the accessibility for blind users
Scope Limitations	The availability and license of JAWS technology
Result	Highlighted areas for improvement
Comment	Significant code amendment required

Test Case	**Trial with LUNAR and ZOOMTEXT Text Magnification User Agent**
Tester	Testing Focus Group
Reason for Test	To evaluate the accessibility for visually impaired users
Scope Limitations	The availability and license of LUNAR and ZOOMTEXT technology
Result	Works well with existing Text Magnification products
Comment	N/A

Test Case	**Trial with WebSpeak User Agent**
Tester	Testing Focus Group
Reason for Test	To evaluate the accessibility for users with dyslexia
Scope Limitations	The availability and license of Webspeak technology
Result	This utility has potential
Comment	Users may require additional training

Test Case	**Configure Content to allow Access to English**
Tester	Testing Focus Group
Reason for Test	To evaluate the accessibility for deaf users
Scope Limitations	N/A

Result	Efficient modification of questions

Test Case	**Trial with Style Switcher**
Tester	Testing Focus Group
Reason for Test	To evaluate the effectiveness of Style Switcher
Scope Limitations	The availability and license of Style Switcher technology
Result	Limited testing
Comment	Exciting area for further development

Note that some 3rd party technical support was required to assist in the setup and testing of the above assistive technologies. Additionally, all of the above assistive technologies require licences.

Observations & Comments

Some of the feedback and observations that were obtained during the testing phase are summarised as follows:

1. Web Site Accessibility

   - The HTML code should conform to WCAG technical standards.
   - Designers need to have a good understanding of how people actually use a site, and the general behaviours of user agents.
   - Content is the third ingredient. How text is written text or label navigation is as much a factor in accessibility as how a developer codes the web page.

2 Assistive Technologies/User Agents

   - It is important to communicate the identity and availability of assistive technologies/user agents in advance. If a visually impaired user uses ZOOMEXT for screen magnification they do not want to be presented with LUNAR. User functionality may be generally the same but commands are different. Similarly if a blind user uses JAWS they may not know how to operate WINVISION.
   - When users are new to a particular assistive technology it is important to have training materials available (on-line preferably) to facilitate new entrants to technology (ICT or AT).

- It is important to be aware that if the user has recently acquired a disability, additional support and sensitivity may be required as the level of vulnerability, isolation and frustration may be greater.
- A period of time may need to be allocated to allow expert assistive technology users configure the user agents to their desired setting e.g. Pitch, Speed and Volume, Voice etc. If setups are altered the default setting should be reset.

3. Web testing automation

   - Many scenarios exist where in general the automated tool validates the code as conforming to standards but the usability of the site is still problematic with user agents. The importance of a diverse focus group works well in ensuring most needs are tested. The user test gives valuable insight into real situations of use.

4. Web page accessibility should be reviewed under the following areas.

Images and image maps	Links	Text (paragraph, horizontal lines, phrases, punctuation and symbols)
Colour usage, coordination	Lists and outlining	Multimedia
User-input forms	Proper use of style sheets (CSS)	Tables
Frames	Use of applets and scripts, blinking, etc	

Accessibility Statement

During the testing phase CFIT made a recommendation to include an accessibility statement; which might state the importance of being open and transparent about accessibility, while inviting difficulties, comments, and suggestions. This facilitates the on-going testing and feedback. A sample accessibility statement might read:

> "Particular care and attention has been made in making this web site accessible. If you come across particular barriers please let us know".

Objective 1: ***How can the design of an accessible system address the problems of Survey Research data?***

Problems with Survey Research Data

Test Case	**What is the likelihood of Missing Data or Values occurring in conjunction with the online survey entry**
Reason for Test	Missing Data or Values can reduce the impact of results to be achieved during data analysis
Testers	Margaret Kinsella
Scope Limitations	Proposed screens at figure 4.4 and 4.5
Result	My Profile – Figure 4.4 The Survey Administration module contains a facility that will allow the survey authors to user define a set of valid values that relates to a specific field. For example, if an Accessibility Category is chosen, a respondent can only select a list of valid entries. This facility exists throughout the application. Survey Entry – Figure 4.5 Questions can be configured as mandatory and therefore respondents must complete the question before advancing.
Comment	Satisfactory results

Test Case	**Eliminate Data Errors from the survey database**
Reason for Test	Check that an acceptable level of data validation is performed
Testers	Margaret Kinsella
Scope Limitations	Survey Administration module Online survey entry module
Result	Data validation has been specified for all data entry fields with the exception of free form text fields e.g. Comments, etc. Some fields contain drop down list, which requires a user to choose a valid value only. Tick box entry has also been enabled. The application has been designed that when an error occurs, no further processing will happen until the error has been corrected.
Comment	Satisfactory result

Test Case	**It should be possible to identify Measurement Errors**
Reason for Test	Sometimes it is necessary to remove data from further processing due to Measurement Errors e.g. wrong question asked. Alternatively, if the data is discovered before it is implemented, it

	should be possible for the survey author to make the changes.
Testers	Margaret Kinsella
Scope Limitations	Survey Administration module
Result	Using a modern database, it is easy to filter bad data using standard SQL facilities. Equally, the Data mining software has standard functionality to easily remove bad data. The proposed survey administration will allow surveys and questions to be changed using standard configuration functionality. Figure 4.8 illustrates how controls or questions are configured. When the survey author amends a control, it will automatically be deployed to the online survey entry module.
Comment	Satisfactory result

Test Case	**Ensure that the data model contains Coding Consistency**
Reason for Test	Check for non-standard units of measurement or value inconsistencies
Testers	Margaret Kinsella
Scope Limitations	Survey Administration module Online survey entry module
Result	The survey application handles Coding Consistency as follows: • Validation will determined by specific hard-coding in the application e.g. the system will check that the valid entries are M for Male and F for Female. This will typically apply when there are a small number of possible values and will likely be handled using radio buttons. • A separate user defined code list will be maintained for some fields where entry requires a drop down list. The list ensures that code consistency will be adhered to.
Comment	Satisfactory result

Test Case	**Identify Bad Metadata**
Reason for Test	Sometimes there is a mismatch between the apparent meaning of a data element and the intended data element definition. For example, this can happen during survey entry.
Testers	Margaret Kinsella
Scope Limitations	Online survey entry module
Result	It is difficult to identify this problem prior to obtaining the results of surveys or uses of the application. It is important that care is taken during the survey design to audit the data prior to deploying the data to the survey entry module. This should involve multiple people reviewing the data in order that any bad metadata is identified up front.
Comment	Satisfactory result

Test Case	**Ensure Sufficient Data Volume is collected**
Reason for Test	A large number of surveys suffer due to insufficient data volumes or too much data. The key aspect of this test is to ensure that the survey application can handle any volume of data
Testers	Margaret Kinsella
Scope Limitations	Survey Administration module Online survey entry module
Result	A modern database i.e. MySQL has been used and is designed to cater for more volumes that would ever be required by the Survey Application.
Comment	Satisfactory result in terms of knowing that the survey application database design and MySQL can handle the required volumes.

Objective 2: ***What are the optimal data mining considerations during the design of the survey framework (incorporating both structured and unstructured data)?***

How can data mining benefit the survey application?

Test Case	**Did data mining identify benefits in relation to scenario 2**
Reason for Test	Data mining defined by Gartner as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques".
Testers	Margaret Kinsella
Scope Limitations	The BUA objectives were focused on data mining to uncover unknown trends in the data only.
Result	The outcomes were inconclusive. A number of clustering algorithms were used; however they did not produce any significant nuggets of knowledge. This is likely to be for the following reasons: • The prepared data that was mined was not of a sufficient volume to generate possible trends • The data collection method was not performed using the survey entry module and therefore had not been properly enriched from the initial survey design. For example, the survey did not have the features that are now available within the new survey application and its ability to assign its many attributes. An interesting observation can be made in relation to the outcome

	of the K-Means algorithm that was performed in both scenario 1 and scenario 2. Figure 4.24 and figure 4.35 illustrates the clusters and their associated groupings that were established for scenario 1 and scenario 2 respectively. The outcome for scenario 2 was more favourable judging by way it determined the grouping which would conclude that the database offered greater potential.
Comment	Inconclusive result. It was difficult to find significant benefits with the dataset that BUA provided. Also, data mining goals were weak due to poor business objectives being set.

Test Case	**Comparison of scenario 1 and scenario 2**
Reason for Test	Both scenarios used the same data but the data was in different formats and data sources.
Testers	Margaret Kinsella
Scope Limitations	The BUA objectives were focused on data mining to uncover unknown trends in the data.
Result	See observations in table 5.1
Comment	The modelling algorithms did not produce any significant results and therefore, it was difficult to compare the scenarios in relations to results. Other comparisons were obvious; such as better data quality using scenario 2; working with a data warehouse in scenario 2 offers greater potential.

Observations	Comment
There were greater data quality issues with scenario 1	For example, sight and hearing contained free form text, where a variety coding inconsistency existed. In addition, a greater number of missing values were identified in scenario 1.
Data mining experts have a preference to working with a data warehouse as was illustrated with scenario 2	This is an observation.
Data preparation in both scenarios focused on identifying missing values, and mapping alphanumeric data to numeric data.	This does not mean that you can assume that data extracted from a data warehouse does not contain missing values, etc
If the data has already been cleansed for a data warehouse then it will most likely will not need further cleaning in order to be data mined. This applied to both scenarios.	N/A
A data warehouse is not a requirement for data mining. However, there are advantages	These advantages have been documented in the evaluation of

Observations	Comment
in having a formal database as per scenario 2.	objective 5 of this chapter
The data warehouse applicable in scenario 2 can help data exploration, for instance by focusing attention on important variables which may not be evident in a flat file or spreadsheet structure.	The survey application database contains additional attributes that were part of upfront data mining considerations. The potential of these attributes were not entirely visible during scenario 2 due to the dataset being used. However, the observation is evident that there is potential for future data mining.
One of the difficulties faced early on was the lack of progress that results when the scope of the business understanding is very broad. After applying various models to the sample dataset, no significant trends were identified. This was the case for both scenarios.	When you have identified the business objectives / requirements they should be defined within the Business Understanding section of the CRISP-DM methodology.

Table 5.1: Observations of two data mining scenarios (author, 2005)

Objective 2: ***What are the optimal data mining considerations during the design of the survey framework (incorporating both structured and unstructured data)?***

How can unstructured data be incorporated into the solution/user interface design?

Test Case	**Can the unstructured data be recorded during an online survey?**
Reason for Test	It is well acknowledged that valuable information is concealed within textual content with no real means of unleashing its potential. Modern data mining software like Clementine, allows data mining and text mining to be analysed together.
Testers	Margaret Kinsella
Scope Limitations	Survey Administration module Online survey entry module
Result	During control or question setup in figure 4.5, the survey author can determine if a separate comment field is enabled during online survey entry. If so, a respondent can choose to quality an answer using a free form text entry. Figure 4.8 shows how the separate 'Comment' is displayed. In addition if a respondent chooses to select 'Other' as a valid

	entry, a separate 'Other – Please Specify' field, as per figure 4.5 can be used to record free form text.
Comment	Satisfactory result

Objective 2: ***What are the optimal data mining considerations during the design of the survey framework (incorporating both structured and unstructured data)?***

What are the optimal data mining considerations during the design of the survey application?

Test Case	**Is the database structure effective for data mining**
Reason for Test	You cannot assume that database design for a specific application will be equally friendly in terms of use by data mining
Testers	Margaret Kinsella
Scope Limitations	MySQL survey application database
Result	In section 4.5.5, containing the data mining process that was implemented using the survey application database, a number of tables were used. In fact the structure made it very straightforward to setup and having such a formal design makes the data understanding job for the data miner a lot simpler. The database design offers other data mining opportunities by way of incorporating some of the other tables that may be relevant.
Comment	Throughout the design, the author was very conscious of the fact of the potential of data mining. Consideration was given to ensuring that additional fields were included for future use and for other survey types. From the outset, the database design is intended to be as generic as possible and is not meant to be limited to learning preferences. For example, three analysis codes have been attached a number of major elements that include the Student/User, Organisation, Goals, Controls, and Survey. Not every survey will have a need for all of them but it adds an extra dimension for data analysis. Another example is the design of accessibility types. Because of the importance of this information to learning preference assessment, the design incorporates multiple types and sub-categories. The emphasis on data analysis certainly focused this requirement and its benefits to data mining. Note that non-learning preferences surveys could find a different use for

	accessibility types. A further example is the facility that allows questions to be excluded from data mining. A separate field on the Control/Question table is used to exclude the bad data from Results processing. This came about as a result of up front data mining considerations.

Test Case	**How can survey configuration assist data mining**
Reason for Test	Poorly asked questions can distort the results that can be achieved by data mining. The problem with paper based surveys is that it is very difficult to change the question when perhaps the questionnaires have already been printed, etc.
Testers	Margaret Kinsella
Scope Limitations	Survey Application Clementine Data mining software
Result	Survey questions are easily re-configured and deployed to the online survey entry module. This works if the survey has not already commenced. In the situation where the survey has been completed, a separate field on the Control/Question table will allow the survey administrator to exclude the relevant question from being included in the results data that will be eventually used by data mining.
Comment	Satisfactory result

Test Case	**Using analysis codes**
Reason for Test	In lots of situations, analysts are interested in studying data by groupings or attributes. This is particularly so when working with large volumes of data. From the beginning of the framework design, the provision of analysis codes was identified as important to providing the necessary groupings.
Testers	Margaret Kinsella
Scope Limitations	Survey Application Clementine Data mining software
Result	A separate set of three Analysis Codes have been applied to the following elements • Organisation • Course • Student / User • Survey • Goals • Controls

	A survey author can designate what an analysis code means to a specific survey and can also user define a set of valid values to ensure coding consistency. Note analysis codes are not mandatory but if relevant can be used to assign attributes to the above elements.
Comment	Satisfactory result

Test Case	**Mandatory entries and Data Entry validation rules**
Reason for Test	One of the biggest problems for data mining is missing values. It has the effect of skewing or distorting results.
Testers	Margaret Kinsella
Scope Limitations	Survey Application Online survey entry module
Result	A number of features have been incorporated into the survey framework design to handle mandatory entries. Every effort has been made to identify all possible areas which may effect the data mining in relation to missing values. A survey author can choose to configure a Control or Question to be mandatory. This allows flexibility to gauge whether a strict regime of mandatory entry is appropriate. A number a data validation rules have been applied to other parts of the application. For drop down lists that require a value, the system will display an error if a non-blank value is not chosen. Where appropriate, a number of hard coding rules have been applied to fields where entry is deemed mandatory.
Comment	Satisfactory result

Test Case	**Effectiveness of Summarising Results**
Reason for Test	Systems such as the Survey application collate large amounts of data in a format that is driven by the requirement of its transactions and master tables. Sometimes this format is not entirely suitable to data analysis tools such as data mining.
Testers	Margaret Kinsella
Scope Limitations	Survey Application Clementine Data mining software
Result	The Survey Application stores the survey recordings in a Header and Detail Table. One of the ways of assessing the data collected is through the use of pre-defined weightings which are typically decided by survey authors, psychologists, etc. Based on the student's answer, a weighting is applied which when all the

	weightings have been accumulated, a measure of success can be determined. The structure of the tables to arrive at a summary measure would make it quite cumbersome for data mining software. Therefore a separate process has been designed to determine weightings and collate the data into a separate results table for use by data analysis tool. Note that this facility has been incorporated into the survey application.
Comment	Satisfactory result

Test Case	**Are historical surveys useful to data mining**
Reason for Test	Sometimes organisations would like to compare the results of one survey to another. Alternatively, organisations might like to combine data from two or more surveys.
Testers	Margaret Kinsella
Scope Limitations	Survey Application Clementine Data mining software
Result	Based on the current experience and understanding of Clementine data mining software, there is nothing technically to prevent this type of analysis being made. This assumes that you have an appropriate way of joining multiple surveys e.g. student/user id. While there is potential to compare historical survey data, it is also important to note that invariably you are not comparing like with like as surveys are generally different. It might be possible that it would work if an organisation repeated the same survey at a different point in time. In the final analysis, the use of multiple surveys can be taken on a case by case basis.
Comment	Satisfactory result

Test Case	**Integrating 3rd party systems with survey application**
Reason for Test	Most organisations operate silos of information, maintaining data about the same individuals in different systems and with different identifications. It would be very useful to combine elements of all related systems to enrich data analysis and enable better decision making. Creating the relationships between entities e.g. students, in data mining software is cumbersome and is best managed as part of the data warehouse.
Testers	Margaret Kinsella
Scope Limitations	Survey Application

	Clementine Data mining software
Result	A separate table has been incorporated into the survey application to maintain 3rd party cross references to the specific entities in the survey application. For example, the user id might have multiple relationships with other systems. In the case of learning preferences this might include systems such as Student Administration, PPS number, Continuous assessment results, etc. Once the relationship mappings have been defined, Clementine data mining software can accept multiple data sources. This is another example where upfront data mining potential enabled such features to be incorporated into the survey application.
Comment	Satisfactory result

5.5 Other Evaluations

Test Case	**Does the survey application provide the basis for a generic survey model?**
Reason for Test	To verify that the survey application provides the basis to configure surveys in addition to the learning preferences assessment.
Testers	Margaret Kinsella in collaboration with NUI, Maynooth who provided a sample survey to test
Scope Limitations	Scope Limitations on implementing full/multiple Survey Research question styles
Result	Overall the acceptance was positive.
Comment	Note a structured survey format is available in Appendix G.

5.6 Chapter Conclusion

The scope of the project had a direct impact on the varied evaluations that were planned and executed. There were many successful tests and evaluations completed. There were also a number of areas identified where further improvements could be made. In the data mining area, it was expected that the results from the two scenarios would have yielded a better outcome in terms of the identification of unknown trends or knowledge that exists in the dataset provided by BUA. The conclusion from this part of the process was that the dataset volume was not sufficient to yield good results. However, the dataset structure and its ability to integrate with 3rd party data sources showed excellant potential.

The accessibility testing proved very interesting and illustrated the possibilities of incorporating assistive technologies into applications. Further conclusions in relation to the research objectives are outlined in chapter 6.

6. Conclusions & Recommendations

6.1 Chapter Summary

This final chapter summarises the key findings from the results in chapter 5, outlines the difficulties encountered, makes suggestions on lessons learned, and highlights interesting observations. This chapter concludes with some recommendations for future work.

6.2 Summary of Key Findings

This thesis report introduced a knowledge discovery framework that encapsulates a set of processes from the initial survey design, to an electronic survey data collection incorporating multiple accessibility features, and concluding with advanced data analysis using modern data mining techniques.

Section 5.4 outlines the specific research objectives evaluations. Each specific objective has been examined and evaluated. The overall conclusion is that the proposed framework can deliver the research objectives.

A key finding during discussions with data mining experts at SPSS highlighted the alarming number of projects that fail. Equally they outlined examples where organisations are having effective outcomes with data mining. These findings not only gave the project a direction in terms of avoiding the common pitfalls but provided guidance in achieving the optimal framework design to maximise the data mining potential.

The evaluation of the research objectives in chapter 5 provides conclusive evidence that up front data mining considerations during the design of the framework solution can have a significant impact.

In summary, the design overcomes many of the problems associated with Survey Research data which as a result has a positive impact on data mining, specifically in the area of Data

Understanding and Data Preparation. Considering that Data Preparation can account for up to 80% of the data mining effort, it is expected that the benefits can be significant.

The formal database design contains several up front data mining considerations. The data warehouse like structure contains a rich set of attributes that can be easily applied across all surveys. For example, a set of user defined analysis codes can be designated for use in one survey and ignored for others. The formal structures will facilitate a better understanding of data for business users and data mining experts which can further optimise the outcomes that can be achieved.

In addition, text mining facilities have been incorporated into the database design as a means of incorporating unstructured data.

The WARP study stated the need for much improvement in web site accessibility. This research outlined a possible accessibility roadmap incorporating the following:

- Web development for a range of user agents, rather than for specific assistive technologies
- Need for change in test practices, to include a more professional test structure in conjunction with accessibility experts and user focus groups. Current practice relies heavily on automated HTML validators which are not sufficient. The web page may meet the WCAG guideline but still remain problematic for the user.
- Incorporation of user preferences with styles switcher to allow change of user interface, font size, colour, contrast, audio etc.
- Separating content from structure with cascading style sheets
- The availability of a wide range of assistive technologies ensuring access to a wide population
- The continued commitment to reviewing web site accessibility, encouraging feedback from user populations, thus fixing it as a permanent agenda item for all web developers.

Difficulties Encountered

In the early stages of the project definition, there were several changes to the scope of the project. Initially the research centred on the development of a Learning Preferences assessment tool incorporating accessibility as a key component. Subsequently the scope changed to include the aspect of Data mining. All of these areas are substantial in their own right and the initial difficulty was limiting the scope in terms of research and development.

In the data mining area, the dataset volume was limited and hence did produce new knowledge.

Lessons Learned

The business understanding phase within the CRISP-DM methodology is critical to defining the goals by which data mining success will be judge upon. For any subsequent data mining project, more time would be given to clarifying very specific goals for the organisation and data mining.

Interesting Observations

While the merging of two different surveys is possible for analysis purposes, there is no likelihood that conclusive results can be generated. Based on the current design, there is a potential that there will be no correlation between two surveys that don't have the same groupings / attributes. Perhaps it is worth investigating whether it is possible to incorporate a survey consolidation facility into the database design.

6.3 Future recommendations

This project has generated much interest and it is exciting to consider the future potential of this project. In order to further enhance the product some possible recommendations are looked at.

If the BUA Learning Preferences survey application were to be available in multiple colleges, the Data Mining possibilities could interpret the data and analyse results from a National perspective.

Incorporate Virtual Learning Environments (i.e. such as Moodle and WebCt) which manage courses, modules, assignments and results. This could have the benefit of analysing student tracking and course participation using Data Mining.

Develop multi-language facilities into the survey application. Dyscovery are based in Wales and have an increasing number of European referrals. Clementine and LexiQuest for Text mining supports multiple language processing.

Similar to the collaborative work with NUI, Maynooth, ITB are recently members of a new Programming research group based in a number of colleges in the UK. The author is the representative for the ITB Programming group. The proposed framework can provide a valuable tool for this forum.

Many of the Retention Initiatives and supports researched during the research discussions have been documented by the author and are under review by the Head of Computing Department to develop a plan specifically for computer science students.

Student profiling is a career-long procedure whereby students develop and maintain a documentary record of their learning experiences. Profiling has been proposed as a way of improving students' ability to take responsibility for their own learning, and of marshalling a portfolio of documented experience that students can use in seeking employment after completing their studies. The application could be enhanced to create industry related outputs. The end result of this project will be a "profiling pack", comprising all the documentation and materials required for the profiling of one student for the length of their academic career.

References

[1] Thatcher, J. & Bohman, P. , *Accessible Web Sites*, Glasshaus, 2002

[2] Dunn et al., *Learning Style Inventory,* Price Systems, 1996

[3] CFIT

[4] *The how's and why's of survey research*, SPSS

[5] Marcell, M. & Falls, A. *Online Data Collection with Special Populations over the World Wide Web*, 2004

[6] Survey Manager from Strategies Group PLC - http://www.strategies.co.uk/products_onlineq.htm

[7] Kirby, A. & Kaplan, BJ. *Specific Learning Difficulties,* Oxford, 2003

[9] Web Content Accessibility Guidelines

[10] Shafer, D. *Designing without Tables using CSS*, Sitepoint, 2004

[11] Higham, N. Designing *TV for Granny, Assistive Technology, Shaping the Future*, IOS Press, 2003

[12] *Data Mining and Knowledge Management in Higher Education –Potential Applications* Jing Luan

[13] Gartner Group - http://www.gartner.com

[14] Osmar R. Zaïane, Ph.D., Associate Professor, Department of Computing Science, University of Alberta - http://www.cs.ualberta.ca/~zaiane/

[15] Rob McCullagh, SPSS

[16] Dr. Larose, D. *Discovering Knowledge in Data*, Wiley & Sons, 2005

[17] Duffin, D. *a*, Centre for Deaf Studies, Trinity College, Dublin, 2004

[18] http://www.dipolar.com

[19] http://www.kitebird.com/articles/access-migrate.html#TOC_3

[20] http://www.caucho.com

[21] http://www.javaworld.com/javaworld/jw-01-2001/jw-0119-jspframe.html

[22] http://www.gulland.com/courses/JavaServerPages/jsp_beans.jsp

[23] University of Washington (http://www.washington.edu/accessit/articles?25

[24] KDNuggets Consulting http://www.kdnuggets.com/polls/2005/data_mining_tools.htm

[25] http://bdn.borland.com/article/0,1410,31863,00.html

[26] Clementine On-line Help

[27] http://www.exxatools.com/SQLion.html

Bibliography

Data Mining

- Han, J. & Kamber, M., *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001
- Pyle, D., *Data Preparation for Data Mining*, Morgan Kauffman, 1999
- Westphal, C. & Blaxton, T., *Data Mining Solutions*, Wiley, 1998
- Various Authors, *Clementine User Manual*, SPSS, 2005

Accessibility and Assistive Technologies

- Craddock, G. & McCormack, L., *Assistive Technology – Shaping the Future*, IOS Press, 2003
- Paciello, M. *Web Accessibility for People with Disabilities*, CMP Books, 2000
- Slatin, J. & Rush, S., *Maximum Accessibility*, Addison-Wesley, 2003
- Thatcher, J. & Bohman, P. , *Accessible Web Sites*, Glasshaus, 2002

User Interface Design

- Lazar, J., *User-Centred Web Development*, Jones & Bartlett, 2001
- Smith, A., *Human Computer Factors, McGraw-Hill, 2000*

Design & UML

- Fowler, M., *UML Distilled*, Pearson, 2004
- Ruble, D., *Practical Analysis & Design for Client/Server & GUI Systems*, Yordon Press Computing Series, 1997
- Sommerville, I., *Software Engineering*, Addison-Wesley, 2001
- Stevens, P., *Using UML*, Pearson, 2000

Web Development

- Barron, D. *The World of Scripting Languages*, Jason Wiley & Sons Ltd, , 2000
- Basham, B. & Sierra, K. & Bates, B., *Servlets & JSP*, O'Reilly, 2004
- Bradenbaugh, J., *JavaScript Application Cookbook,* O'Reilly, 1999
- Castro, E. *HTML for the World Wide Web*, Peachpit Press, 2000
- Lowery, J. *Dreamweaver MX*, New Riders, 2004
- Patzer, A. *JSP Examples and Best Practices*, Apress, 2002
- Various Authors, *Professional JSP*, Wrox, 2000

Database

- DuBoise, P., *MySQL Cookbook*, O'Reilly, 2003
- Reese, G. & Yarger, R., *Managing and Using MySQL*, O'Reilly, 2002

Learning Styles

- Biggs, J. *Student Approaches to Learning and Studying*, Acer, 1987
- Harri-Augstein, S. & Thomas L. *Learning Conversations*, Routledge, 1991
- Shrey, D. and Lacerte, D. *Principles and Practices of Disability Management*, CRC Press, 1997
- Fisher, D., Sax, C., Pumpian, I., *Inclusive Learning from Contemporary Classrooms*, Paul Brookes Publishing Co, 1999

Research

- Cohen, L. and & Manion, L. *Research Methods*, Routledge, 1994
- Roberson, S. & Robertson, J., *Mastering the Requirements Process*, Addison-Wesley, 1999