HOW AN AGILE DATA WAREHOUSE METHODOLOGY DESIGNED BY A UNIVERSITY REPORTING TEAM IS ENABLING FLEXIBLE AND RAPID ANALYSIS

Kristina Kaulenas

ABSTRACT

The Australian higher education sector is undergoing a period of substantial change. This has led to an increasing focus on metrics, targets and evidence based decision making. For instance, the Government has set a focussed agenda on improving university participation and quality. Institutions will need to negotiate performance based funding targets as part of their mission based compact agreements, and the launch of the Higher Education Participation and Partnerships Program (HEPPP) has raised the profile of equity measures. Similarly, the Cycle 2 of the AUQA audit, with its focus on benchmarking and academic standards, has highlighted the need for accurate data on student academic performance and graduate attributes.

If universities are to respond to these agendas they will need access to accurate and timely data to support institutional research and management decision making. The ability to react quickly to changes in government methodology and analytical requirements of institutional researchers and senior management, as well as to respond to AUQA's audit data requirements, will be a key function of university statistics units.

In 2006, the Office of Planning and Quality (OPQ) at Monash University commenced work on the design and development of a data warehouse focussed on supporting DEEWR reporting, institutional research and general information dissemination. The development of this data warehouse differs from many others. It has been developed entirely in house by analysts who are subject matter experts and are involved in all aspects of the project – from data extraction, cleansing and preparation, through to all data warehouse development and maintenance activities, metadata management and data dissemination.

This paper will provide an overview of the structure and content of the OPQ data warehouse as well as examples on how its development has improved the provision of timely analysis to support evidence based decision making.

BACKGROUND

The Australian higher education sector is undergoing significant change. Recommendations from the Bradley Review of Higher Education, the launch of the Higher Education Participation and Partnerships Program (HEPPP) and the upcoming Cycle 2 of the Australian University Quality Agency (AUQA) audit at Monash University in 2012, only increase the role played by university planning and statistics areas to provide accurate and timely data for evidence based decision making.

Additionally the Government will introduce significant amounts of performance based government funding associated with various agreements made between each university and the federal government. The

establishment of these 'mission based compacts' allow universities to pursue their own distinct strategic direction whilst at the same time contribute to the Government's sector wide higher education objectives (DEEWR, 2009a). The provision of public funding necessitates transparency and accountability and the ability to meet institution specific performance targets. Universities are highly accountable to government and accrediting agencies and have dramatically different reporting requirements and legislative obligations than private sector (Nakatani & Chuang 2005).

In a typical university, the planning and statistics areas are the core areas responsible for co-ordinating and ensuring adherence to the university's strategic planning framework and managing the reporting and analysis of some or all of the DEEWR data collections. Depending on their data management framework and capabilities, the services that their analysts provide will range from standard and ad-hoc reporting through to scenario planning and predictive modelling. The department is also usually responsible for monitoring the University's performance and generally providing research to improve strategic decision-making and to ensure that the University meets its legislative goals and targets.

So how will external changes in funding and auditing affect the operation of university planning and statistics departments? One of the main focuses of the Cycle 2 AUQA audit will be academic standards and the auditors will be looking for systematic tracking of the performance of student cohorts including admission, success, progression and completion as well as their graduate satisfaction (AUQA 2010). This monitoring will have to come from the core planning and statistics areas within the university who are experts in the interpretation and analysis of this data.

Similarly, whenever there is funding attached to performance indicators there is a requirement to replicate published figures and to analyze the data at a finer level of detail in order to elicit where performance is good and where improvements can be made. Sometimes, replicating these figures can be a complex and time consuming process. The indicators used to allocate HEPPP funding are a prime example.

The HEPPP will encourage participation of students from low socio-economic (SES) backgrounds in higher education and universities will share in more than $56 million to be allocated in 2010, increasing to $168 million in 2013 (DEEWR, 2010a). The performance of universities in regard to the recruitment of low SES students is calculated using a formula which takes into account the home location of the student as well as receipt of any income support payments.

The work involved in determining the socio-economic status of a student's permanent home location alone will require the determination of the census collection district (CCD) of a student's permanent home location address. This in itself will require significant data coding to transform and clean student addresses in order to prepare them for geocoding. The Centrelink component of the formula is sourced from within DEEWR and is only currently available at the aggregate institutional level. Being able to replicate the overall figure for Monash University and then break this down by faculty/campus has already been raised as information that is of vital importance to the university – to assist in the determination of strategy for increasing this cohort of students within the university.

Given the increasing focus on monitoring of academic standards and evidence based decision making, it is vital universities find ways to manage and store their data to enable flexible, rapid yet accurate analysis. Universities hold vast amounts of complex information so this can seem like an overwhelming task. To tackle this, in late 2005 Monash University finalised an enterprise-wide information management strategy. A subsequent business intelligence (BI) strategy was developed during 2006 to establish a data warehouse and BI infrastructure in order to merge the numerous information repositories into a comprehensive and robust framework. The aim was to

enable the provision of data to support managerial decision making and to help monitor performance of key university performance indicators. KPI reporting was seen as the pinnacle of reporting from the proposed environment (Monash, 2006a).

It was also around this time that the University Planning and Statistics (UPS) team within OPQ began to assess their own data management framework. The university wide BI project was still in its infancy and UPS was already managing numerous datasets from the DEEWR student and staff data collections (both Monash and national benchmarking data) as well as a range of VTAC admissions and student survey data.

This paper will outline the history and development of the UPS data warehouse as well as provide an overview of its structure, contents and capabilities. It will also discuss the reengineering of the overall data management framework within UPS, important lessons that have been learnt along the way, as well as future implementation plans for the warehouse. It is will also discuss the specific developmental methodology utilized in building the warehouse and consideration will be given to the suitability of the standard systems development life cycle (SDLC) on BI projects - due to the difficulty in accurately determining exact requirements at the beginning of such projects (Behrangi et al, 2007).

Monash reporting environment 2006

The availability and production of management information within the university in 2006 could be considered immature on the business intelligence maturity spectrum and was characterised by the following (Monash, 2006a):

1. Inconsistent views of data across the university.
2. Manual data collation and manipulation.
3. Limited analytical capabilities.
4. Multiple operational systems.
5. Data quality issues.
6. Growth and complexity of information at Monash was increasing exponentially each year – more sources of data being stored in increasingly complex structures with little overall communication between the various systems – poorly integrated IT architecture designed to support the day-to-day operations of distinct operational transactions (Gibson & Arnott, 2010).

UPS could be seen to be a microcosm of this. As one of the most advanced data users and data custodians within the university, the problems associated with data and procedural inconsistency, redundancy and quality were all present – albeit on a smaller scale.

In 2006 UPS was utilizing a series of historical SPSS flat files – with associated value and variable labels. These files were created using SPSS code primarily for the purposes of statistical analysis within SPSS where everything required for analysis was located within a single file.

There were many advantages to this sort of file structure for institutional research and it was a big improvement over the data management practices within the unit prior to its introduction. The graphical user interface and the presence of command syntax enable reproducibility of repetitive tasks and procedural documentation (the syntax preserves a record of the data preparation and analysis processes). Unlike an Excel spreadsheet, the 'variable view' within SPSS incorporates a metadata dictionary (data about data) which is saved within the data file.

Monash has a site-wide license for SPSS and it is widely used by researchers and public and private sector business professionals for its statistical functions, easily transportable output and ease of use. SPSS code is easy to write and interpret for people with little or no programming experience and the menu interface can also be used to access almost any statistical analysis or data management function within SPSS. With the appropriate configuration, SPSS has advanced data import/export functionalities and ODBC connectivity.

Whilst the introduction of SPSS had many advantages (and it is still in use today within UPS for its statistical analysis capabilities), as the information requirements became more complex and the quantity of data maintained within UPS grew, the responsibilities for the management of the suite of SPSS code was dispersed amongst several UPS staff members. It became clear that there was a growing need for a co-ordinated information management strategy and the centralisation of the business logic and the data itself.

Below is a list of some of the issues that were becoming more evident as time went on:

- Same variable had different variable or value labels within different historical files.
- No common naming conventions in place.
- Difficult to determine the role a variable was playing in a particular file or how it was derived.

- Code redundancy made change control difficult. Due to the inter-dependencies between the files a staff member making a change required systems view knowledge of all business processes within UPS.
- SAS (another statistical programming language) was being used by some staff members and SPSS was being used by others (for data preparation and management).
- Multiple versions of the same file in various physical locations.
- Manual and error-prone reporting of thousands of static web reports. Majority of staff time spent of data preparation and then copying and pasting data from pivot tables into static report templates.
- Staff members managing data management and preparation processes had varying levels of technical proficiency with no over-arching commonality in their approaches.

To manage these data management and quality issues OPQ, in 2008, commenced a journey down the dimensionally modelled path and began development of a data warehouse.

Flat files, relational databases and dimensionally modelled data warehouses:

To understand the benefits of a dimensionally modelled data warehouse, it is first important to understand the differences between a flat file database and a relational database. A flat file database is one that is designed around a single physical table. One which contains all the fields required for a specific purpose – often with duplicate data across multiple columns and records (see figure below which demonstrates the fields in a sample historical SPSS flat file).

Name	Type	Width	Decimals	Label	Values
record	Numeric	1	0	E300 Record type	{1, Not reported}...
year	Numeric	4	0	E550 Reference Year	None
id	Numeric	8	0	E313 Student ID	None
course	String	10	0	E307 Course code	None
coursex	String	10	0	E307 Original course code in Callista	None
CRS_ABBREV	String	25	0	Callista latest version abbrev	None
CRS_NAME	String	100	0	Callista latest version title	None
crs_grp	Numeric	1	0	Course type group	{1, HD Research}...
CRS_TYPE	Numeric	2	0	E310 Course of study type code	{1, Higher Doctorate}...
COMBINED	String	2	0	E455 Combined course indicator	{0, Not a combined course}...
FOE	Numeric	6	0	E461 Field of education	None
FOE_NAME	String	80	0	E461 Field of education name	None
FOE_BROAD	String	46	0	E461 Field of education broad categories (2 digit)	None
FOE_NARROW	String	60	0	E461 Field of education narrow categories (4 digit)	None
FOE_SUPP	Numeric	6	0	E462 Field of study supplementary code	None
FOE_SUPP_NAME	String	80	0	E462 Field of education name	None
FOE_SUPP_BROAD	String	46	0	E462 Field of education broad categories (2 digit)	None
FOE_SUPP_NARROW	String	60	0	E462 Field of education narrow categories (4 digit)	None
MFACULTY	String	15	0	Managing faculty	None
DFACULTY	String	16	0	Degree faculty	None
S_LEAVER	Numeric	8	2	E925 School-leaver indicator	{.0, Overseas student}...
ADMIT_OLD	Numeric	2	0	E327 Basis of admission pre-2005	{0, Missing}...
ADMIT_NEW	Numeric	8	2	E327 Basis of admission 2005+	{1.00, Not a commencing student}...
C_BIRTH	Numeric	4	0	E346 Country of birth code	None
C_BIRTH_NAME	String	36	0	E346 Country of birth country name	None
C_BIRTH_SUBREGION	String	28	0	E346 Country of birth subregion	None
E319	String	5	0	E319 Term residence	None
E319_AREAS	String	28	0	E319 Term residence area	None
E319_COUNTRY	String	36	0	E319 Term residence country name	None
E319_OSREGION	String	28	0	E319 Term residence overseas subregion	None

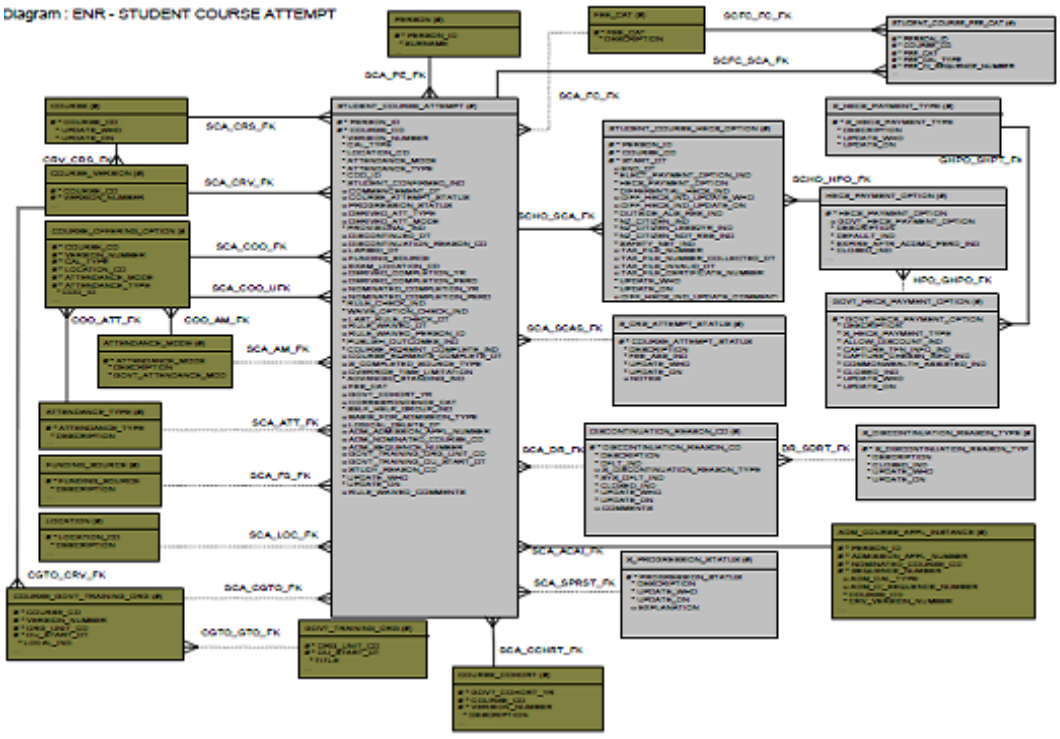
Designing flat files is simple and requires little database design knowledge. By comparison, a relational database is a collection of related tables and uses a mathematical theory called normalisation to model these relationships in

order to minimise data duplication, reduce data anomalies and to reinforce data integrity – an essential requirement of transaction based processing (Kimball et al, 2002) – such as those within a student administration system.

For example, while a flat file of student enrolments might store the course code, course name, course abbreviation, course type etc. for every student, a relational database will store the course code for every student within the main table and have a course reference table containing the course code and related course characteristics which can then be linked back to the student record. In reality all the characteristics associated with a course may physically be located in dozens of reference tables in a relational system (e.g. course names may be stored in one table while course fields of education will be stored in another and course types within yet another). Designing a relational database therefore requires more planning to fully describe how the data is organised in terms of data structure, integrity, querying, manipulation and storage.

The downside of a relational structure is that we are often left with hundreds (or even thousands) of tables linked together in a bewildering web of relationships – conceptually represented by database developers by numerous entity-relationship diagrams (ERD's). Most business users cannot intuitively understand or navigate such a complex model, plus data retrieval speeds are slow, and complex analysis often results in performance degradation of operational systems.

Below is an example of a small portion of an ERD from Callista (Monash's student system) focused on a 'student course attempt' transaction.



By comparison, data warehouses are designed for fast and intuitive retrieval of integrated and high quality data for reporting and analysis. Like relational databases they also contain multiple tables that are related to each other,

but rather than the focus being on minimising data duplication, they are structured in a way that reflects how a business operates – through the establishment of 'subject-oriented' central fact tables (e.g. a table containing student enrolment records) and associated dimension tables (e.g. a course dimension table containing all course characteristics rather than storing different course characteristics in different tables) – which are 'conformed' to enable them to be shared across different subject areas (e.g. the course dimension table is shared across the Monash enrolment, load and completions data and also used in some admissions and student survey views).

Enhanced Data Dissemination Project (EDDP) 2006 - 2008

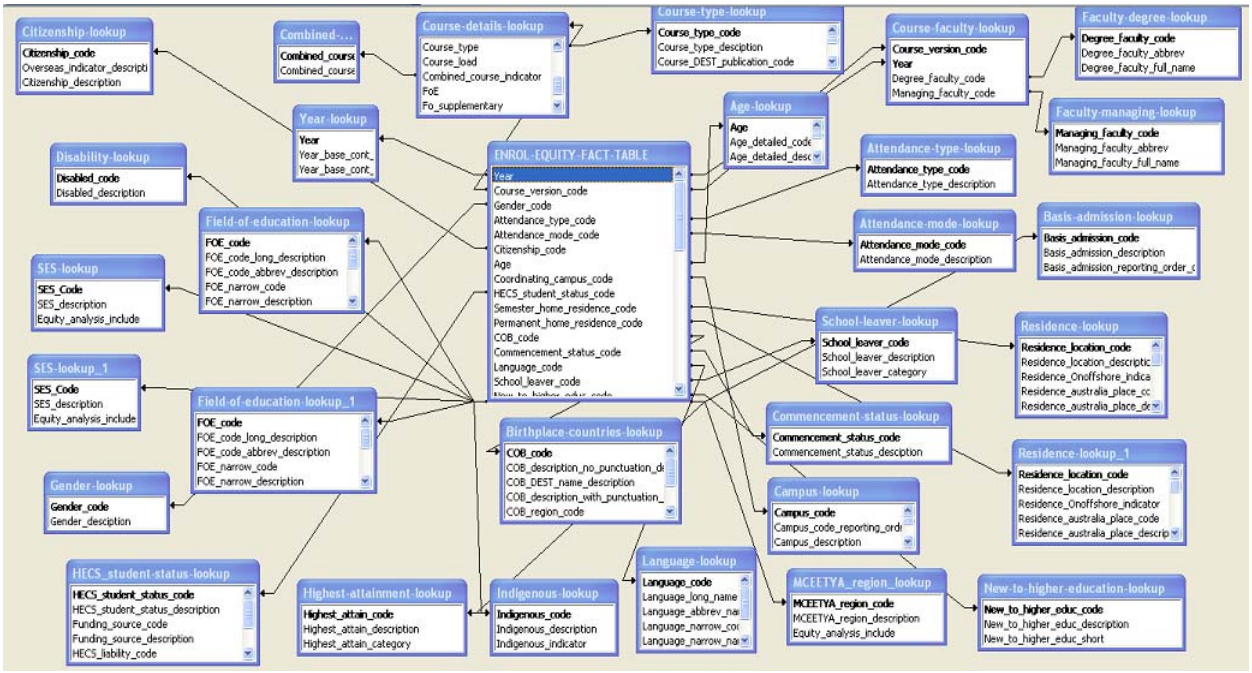
The requirement for a central repository/data warehouse of UPS data was first officially identified in September 2006 in a '2007 Education Related Strategic Initiatives application'. The proposed project was to provide easier access to more data that would support the Education Plan but primarily to support the objective of achieving "consistently superior results in indices, rankings, competitions, external audits and assessments" (Monash, 2007 p.3).

In mid-2006 another department within Monash launched a pilot project which was to include data in a Cognos BI 8 (Metrics Manager) presentation layer. UPS also followed suite around the same time on a 'KPI Reporting project' which was to be a two-staged implementation. The scope was to commence with a limited set of education indicators and then in Stage 2 progress to across the board Education, Research and Business Management KPI's, including rankings and drill-down capability through the Cognos Metrics Manager Scorecarding functionality (Monash, 2007).

As this project progressed it became apparent that the flat file structures that predominated UPS would not support the aims of the project. In early 2008, UPS engaged the services of two Cognos consultants to develop some enrolment and equity reports in Cognos Report Studio. The development of these reports highlighted the problems associated with the inconsistencies in the UPS data.

This was a pivotal stage in the development of the EDDP as the focus of the project shifted from the implementation of web-based reporting tools to the establishment of a data warehouse focused on supporting the reporting and analysis activities within UPS. One of the UPS project team members embarked upon developing a pilot Access data warehouse which was limited in scope but contained all the elements required to build the Cognos reports. The project team member was able to develop a single fact table and associated dimension tables in a period of 2-3 days (see below).

While this improved the integrity of the data, it did not solve all of the reporting issues as the Cognos software could not connect directly to an Access database. Therefore an investigation was launched into establishing a UPS database schema in the Callista Oracle environment to enable Cognos to link directly to Oracle and to also take advantage of the Callista security framework.



The actual development of a limited scope Oracle data warehouse provided an opportunity for the project team and UPS management to see first-hand how a centralized data repository would benefit the department and resolve many of the data quality and data management issues.

It took the development of a pilot deployment of the UPS schema in Oracle to prove the concept of the data warehouse and get management and even project team buy-in. The analysts within UPS had to be convinced that they could still get access to the data that they would require without the presence of the SPSS historical flat files and management had to be convinced that this was achievable with the limited resources available within UPS.

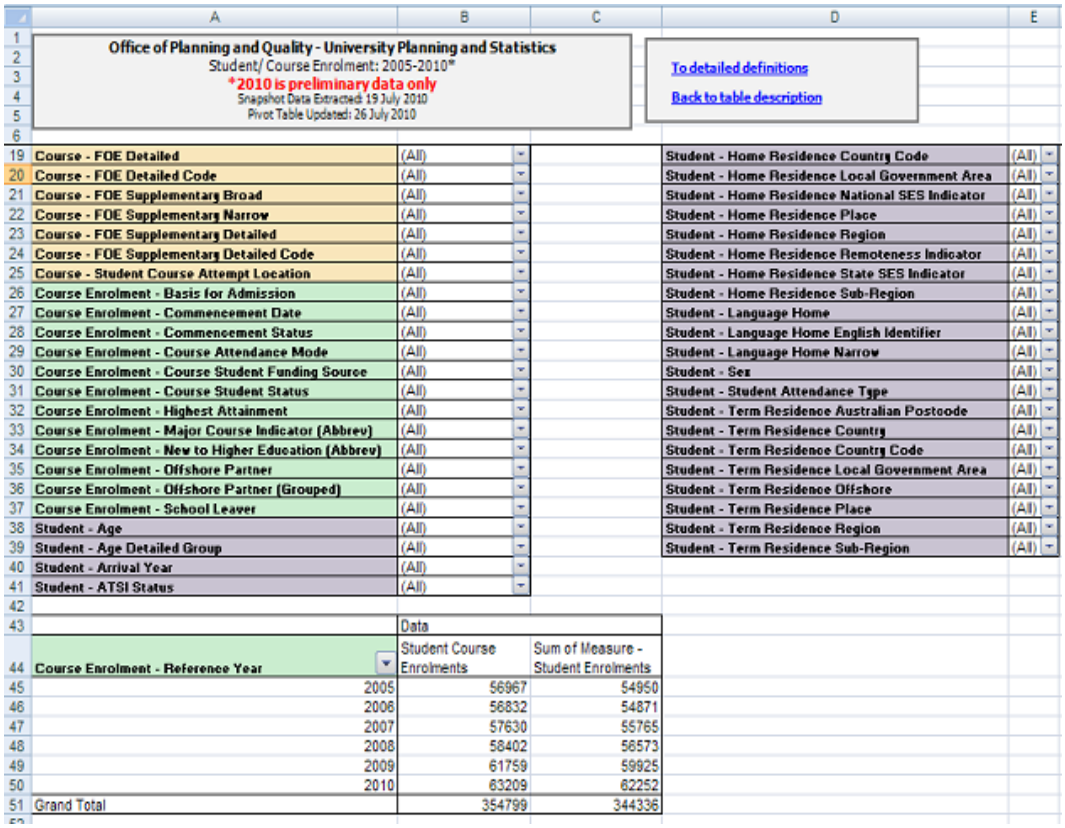
Oracle warehouse development

Despite no one in the project team having any Oracle development experience (other than a basic understanding of SQL as a database querying language), and no central IT support to migrate the data to an Oracle environment , the setup of a UPS schema within the Oracle instance occurred within a matter of days.

UPS was able to migrate the data into the Oracle schema within 2-3 days. What followed over the next couple of years was an incremental development of the UPS Oracle data warehouse - with all developmental activities being carried out and managed by UPS staff. Instead of neatly delineating the developmental activities, necessity demanded that the Systems Manager within UPS become cross-trained in all facets of data warehouse activities. This included:

- All Oracle developmental activities involved with building the data warehouse schema.
- All ETL programming and data quality activities including resolution of any definitional and methodological issues.

- All conceptualisation and project management activities associated with defining the overall data warehouse architecture, including establishment of naming conventions (and ensuring adherence to them.
- Change control management across DEV, QA and PROD environments.
- Development of role playing dimensions and materialized views to support information dissemination and analytic requirements.
- Information dissemination via suite of custom UPS pivot tables configured to connect to the UPS data warehouse schema via ODBC Oracle Client.



Office of Planning and Quality - University Planning and Statistics
Student/ Course Enrolment: 2005-2010*
*2010 is preliminary data only
Snapshot Data Extracted: 19 July 2010
Pivot Table Updated: 26 July 2010

To detailed definitions
Back to table description

Course - FOE Detailed	(All)			Student - Home Residence Country Code	(All)
Course - FOE Detailed Code	(All)			Student - Home Residence Local Government Area	(All)
Course - FOE Supplementary Broad	(All)			Student - Home Residence National SES Indicator	(All)
Course - FOE Supplementary Narrow	(All)			Student - Home Residence Place	(All)
Course - FOE Supplementary Detailed	(All)			Student - Home Residence Region	(All)
Course - FOE Supplementary Detailed Code	(All)			Student - Home Residence Remoteness Indicator	(All)
Course - Student Course Attempt Location	(All)			Student - Home Residence State SES Indicator	(All)
Course Enrolment - Basis for Admission	(All)			Student - Home Residence Sub-Region	(All)
Course Enrolment - Commencement Date	(All)			Student - Language Home	(All)
Course Enrolment - Commencement Status	(All)			Student - Language Home English Identifier	(All)
Course Enrolment - Course Attendance Mode	(All)			Student - Language Home Narrow	(All)
Course Enrolment - Course Student Funding Source	(All)			Student - Sex	(All)
Course Enrolment - Course Student Status	(All)			Student - Student Attendance Type	(All)
Course Enrolment - Highest Attainment	(All)			Student - Term Residence Australian Postcode	(All)
Course Enrolment - Major Course Indicator (Abbrev)	(All)			Student - Term Residence Country	(All)
Course Enrolment - New to Higher Education (Abbrev)	(All)			Student - Term Residence Country Code	(All)
Course Enrolment - Offshore Partner	(All)			Student - Term Residence Local Government Area	(All)
Course Enrolment - Offshore Partner (Grouped)	(All)			Student - Term Residence Offshore	(All)
Course Enrolment - School Leaver	(All)			Student - Term Residence Place	(All)
Student - Age	(All)			Student - Term Residence Region	(All)
Student - Age Detailed Group	(All)			Student - Term Residence Sub-Region	(All)
Student - Arrival Year	(All)				
Student - ATSI Status	(All)				

	Data	
Course Enrolment - Reference Year	Student Course Enrolments	Sum of Measure - Student Enrolments
2005	56967	54950
2006	56832	54871
2007	57630	55765
2008	58402	56573
2009	61759	59925
2010	63209	62252
Grand Total	354799	344336

- Establishment and management of data warehouse metadata.

The UPS data warehouse could be considered a non-standard data warehouse implementation due to the minimal involvement from central IT and the number of specialised data warehouse team roles assumed by the UPS Systems Manager. While other staff members within both the reporting and analysis teams within UPS were drawn upon to provide advice and assistance, the UPS Systems manager played the role of project manager, data warehouse architect, database developer, business analyst, data specialist, ETL modeler and developer, information delivery manager and end-user support and training.

While having one staff member perform so many roles is not usually best practice, it did have the benefit of allowing the project to remain agile. It was not constrained by a cross-functional, multi-level governance structure or a structured project management approach. The focus of the project was on delivering a working data

warehouse that could continue to support the analysis and information dissemination requirements of UPS throughout all stages of its implementation through the development of a robust and centralized information management strategy and framework.

The current state of the data warehouse

Unlike an enterprise data warehouse which aims to consolidate data for the whole organization (Kimball et al, 2002), the UPS data warehouse could be considered a 'datamart' – as it was limited in scope to satisfy the reporting requirements of its analysts and well as those involved in the support of the university's strategic planning framework and university wide reporting and monitoring.

The data warehouse currently contains the following modules:

- Monash DEEWR student and staff collection census data (enrolment, load-liability, completions and staff - including preliminary 2010 data updated monthly). This enables enrolment, load, equity, academic performance (GPA, Average marks, Grade Distribution, Student Course Progression, Retention Rate), Student-Staff Ratio, Course completion etc types of analysis.
- National aggregated DEEWR student and staff collection data. This enables comparison of Monash student and staff data against other institutions nationally.
- Go8 record-level student collection data. This enables student tracking across years and the replication of various performance indicators and measures and the benchmarking of Monash data against the Go8 for KPI reporting.
- VTAC admissions data – sector wide data on applicants, preferences, VCE subject outcomes, Popularity Polls (course preferences) and Graduate Entry Teaching data. This enables analysis at the applicant or preference level of detail.
- National Graduate Destination Survey (GDS), Course Experience Questionnaire (CEQ) and Postgraduate Research Experience Questionnaire (PREQ) data. This enables analysis of full-time employment, further study, CEQ scales/survey items, graduate salaries, PREQ scales etc.

Along with the core DEEWR submission data, additional reporting elements are extracted from Callista to enable Monash student performance analysis. UPS disseminate data to the University community through a series of customized Excel pivot tables. Most of these now have direct ODBC links to the data warehouse – so they can be refreshed automatically.

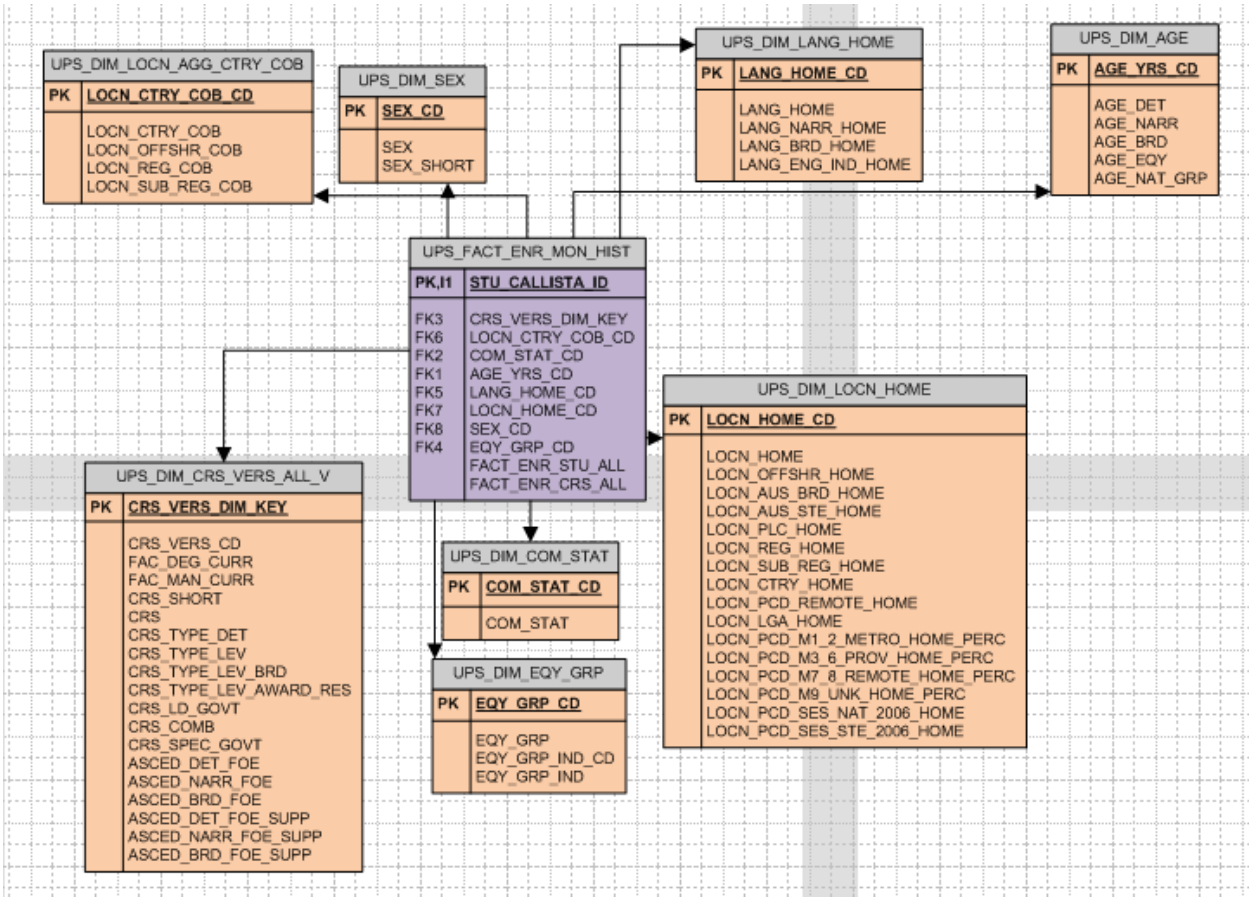
Examples of how the data warehouse enables rapid and flexible analysis

How does the data warehouse enable UPS to respond rapidly to additional reporting requirements? To best illustrate this, the section below will focus on a single module within the overall data warehouse and how the materialized view functionality is used to create views within the data warehouse that can then be used for analytical or data dissemination (i.e. pivot table) purposes.

The unit's equity data is created from the enrolment and load fact tables and associated dimension tables. Below it will also discuss how the view could be extended in order to enable replication of the new HEPPP participation formula for the determination of low SES students.

The simplest structure of a data warehouse is called a star schema – with a central fact table and associated dimension tables (which contain descriptive textual information and the attributes which are actually used for analysis and reporting).

In the diagram below, the fact table holds the data on Monash 'student course enrolments' and the following columns about this subject area within the warehouse: course of enrolment and its commencement status and the following student demographic characteristics: country of birth, age, language spoken at home, permanent home location, gender and whether the student is used in the calculation of the equity measures. The fact table also has two 'facts' (measures): student count (FACT_ENR_STU_ALL) and course enrolment count (FACT_ENR_CRS_ALL).



The associated dimension tables include: UPS_DIM_LANG_HOME (containing language fields), UPS_DIM_AGE (containing age groupings), UPS_DIM_LOCN_HOME (containing home residence groupings based on postcode or country) and UPS_DIM_CRS_VERS_ALL_V (containing course characteristics), and they each describe the codes within the central fact table at various levels of detail. Dimension tables enable users to 'slice and dice' data in the data warehouse and have three primary functions – to provide filtering, grouping and labeling (Wikipedia, 2010).

For example – the base dimension table 'UPS_DIM_LOCN' – at its lowest level of detail can describe the individual Australian postcode or overseas country code, however also has attributes which enable this characteristic to be

grouped at the Australian state or country region and sub-region level of detail, as well as indicate the socio-economic status of the postcode and its metropolitan/provincial or remote classification– see below.

LOCN_CD	LOCN	LOCN_OFFSHR	LOCN_AUS_BRD	LOCN_AUS_STE	LOCN_REG	LOCN_SUB_REG	LOCN_CTRY	LOCN_PCD_M1_2_METRO_PERC	LOCN_PCD_M3_6_PROV_PERC	LOCN_PCD_M7_8_REMOTE_PERC	LOCN_PCD_SES_NAT_2006
A2748	2748 Orchard Hills	Australia	Interstate	New South Wales	Australia	Australia	Australia	1	0	0	Medium
A3337	3337 Melton	Australia	Melbourne	Victoria	Australia	Australia	Australia	0.9492	0.0508	0	Low
A3211	3211 Little River	Australia	Rest of Victoria	Victoria	Australia	Australia	Australia	0	1	0	Medium
X1201	New Zealand	Overseas	Overseas	Overseas	Other Oceania and Antarctica	New Zealand	New Zealand	999	999	999	No information
X2401	Denmark	Overseas	Overseas	Overseas	North-West Europe	Northern Europe	Denmark	999	999	999	No information
X5104	Thailand	Overseas	Overseas	Overseas	South-East Asia	Mainland South-East Asia	Thailand	999	999	999	No information
X8102	Canada	Overseas	Overseas	Overseas	Americas	Northern America	Canada	999	999	999	No information

Dimension tables are easily extendable to allow for the inclusion of new attributes. For example, if there was a requirement to group the Australian postcodes in a very specific way for reporting of metropolitan and regional areas – this could be accomplished through the creation of a new category within UPS_DIM_LOCN and the assigning of individual postcodes to this new category.

In any data warehouse there will be multiple fact tables, and dimension tables need to be shared across the warehouse. Dimension tables can also be used for numerous roles within the warehouse i.e. the base location dimension (UPS_DIM_LOCN) – plays the following roles within the warehouse: permanent home location, semester location and commencing location. This base dimension may also need to be summarized to provide a more aggregated version of the data i.e. to describe a student or staff member's country of birth.

A series of dynamic materialized views of this base dimension table enable it to be used in multiple ways across the entire warehouse depending on its context and to ensure consistency in how the data is presented for analysis or reporting. The importance of naming conventions also comes into play to describe the role that the variable is playing as well as its relationship to the base dimension.

In order to enable the replication of the participation component of the HEPPP fund – this would require the integration of data relating to a student's SES (based on the CCD of their permanent home address) as well as their entitlement to Centrelink benefits. In order to enable this, processes would have to be designed to firstly determine the appropriate student address to use for a particular student, and then subsequent geocoding of the address would provide a CCD which could then be matched to a SES. Student Centrelink data would also have to be incorporated into the warehouse with some way of identifying whether a student is entitled to Centrelink benefits at a particular point in time.

In its most simplistic implementation, the above star schema would simply join to two additional dimension tables. The first new dimension table would contain student address information, CCD (if able to determine) and associated SES details for that address. The second dimension table would contain student Centrelink entitlement data at particular points in time. Both dimension tables could be matched to the enrolment data by matching on the Student ID and/or Year. The resulting analysis view would then contain the SQL required to consolidate the variables from the various dimension tables associated with the particular fact table/s, and perform any additional transformations or calculations to produce a final data view which has all the variables required for analysis.

Currently a single equity analysis view enables the determination of access, participation, retention and progression rates and ratios for each of the equity groups that DEEWR publish in their IAF report.

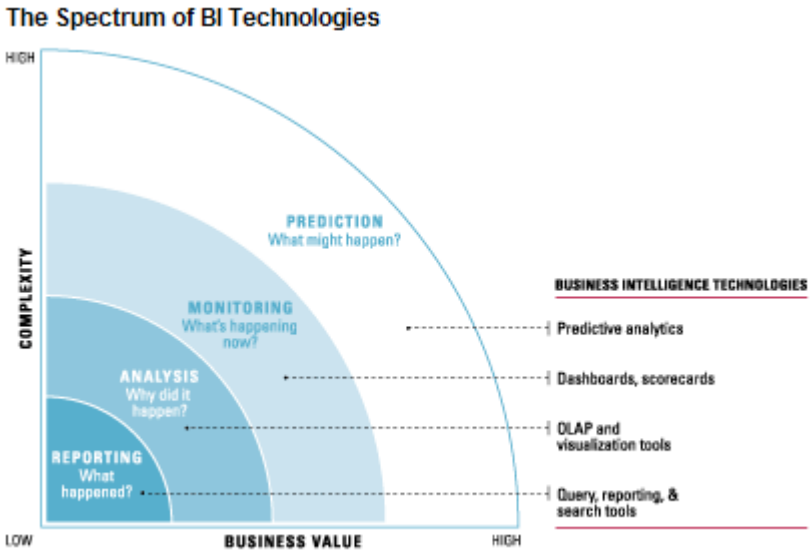
A revised equity view would theoretically contain a couple of new calculated facts for a student enrolment record, which would represent the 'A' and 'B' component of the HEPPP participation fund formula below (DEEWR, 2010a, p. 15):

$$C = (2 \times A + B) / 3$$

- Where '**C'** = Indicator of undergraduates from low SES backgrounds. This would be calculated dynamically within the pivot table based on the values of 'A' and 'B' – which are recorded at the student enrolment level, but could be calculated at any level of aggregation of the data.
- **A** = record in equity analysis view will have a value of 1 in this variable if the student is a low SES student – based on their cleaned and geo-coded permanent home address.
  B = record in equity analysis view will have a value of 1 if they are entitled to Centrelink benefits at a particular point in time.

The replication of the Centrelink component of the HEPPP participation formula will be methodologically difficult as DEEWR only obtain this data at the aggregate institutional level for use within their calculations; however UPS are currently investigating the possibility of sourcing this data. Further work is still required in developing the methodology and incorporating this into the warehouse to allow for replication of this data at lower levels of detail.

The foray into geocoding of student addresses will pave the way for more sophisticated analytical techniques associated with 'location intelligence'. Some analysis is best done using a visual/geographical representation of the data, and the future integration of further ABS data into the warehouse will enable analysts to have access to a much richer range of data to aid any investigative needs and to progress further along the business intelligence spectrum (see diagram below) – from simply reporting what has happened in the past to predicting what might happen in the future.



Future implementation plans

Whilst UPS has come a long way since the beginning of 2008 with respect to their data management framework, there is a need to further consolidate the currently disparate SPSS, SAS and SQL processes into an integrated solution, and to take advantage of other capabilities that a product such as Oracle Warehouse Builder (OWB) might offer, such as: data lineage, data mapping, impact analysis, dimensional modeling, data quality and documentation features.

For example, the impact analysis feature of OWB would enable UPS to determine the effects throughout the entire warehouse of a change to a particular variable or data source, and the data lineage feature enables tracking the path that any variable takes throughout the warehouse and any transformations that occur to it. Currently this is possible but time-consuming and complex.

Monash University has also been using up until 2010, Oracle Discoverer as its university-wide BI presentation/reporting tool. It was initially chosen as a low-cost solution to provide experience with BI technology and to give an opportunity to further develop user requirements (Monash, 2006a). In 2010 Monash University chose SAP Business Objects as its BI presentation layer. OPQ are currently working with the Monash BI team to investigate synergies between the enterprise data warehouse and the OPQ 'datamart' and will be reviewing their information dissemination and analysis strategies over the next couple of years.

Conclusion

The benefits associated with having a centralized data repository have resulted in a much closer collaboration between the reporting and analysis teams within UPS, a greatly reduced effort in preparing and disseminating high quality and consistent data, as well as the ability to rapidly respond to changing reporting and analytical requirements.

The analysis team is now able to spend much more time analyzing the data and performing more complicated forms of analysis – such as scenario planning, preparation of issues papers, student tracking/cohort analysis and the preparation of custom datasets and pivot tables. For instance, the consistency across the different modules allows analysts to easily merge together data on different subjects areas with ease (e.g. examine how many deferred students return by matching together VTAC admissions data containing deferral records with subsequent enrolment data).

The unique history and developmental approach within UPS mean that this is a flexible data warehouse designed by analysts for analysts and business users. The iterative approach to development has meant that it could rapidly gain management and user support and provide real benefits to the department within a very short period of time.

The production of a pilot data warehouse in early 2008, which was limited in scope but was able to successfully demonstrate the concept of data warehousing, was critical in the success of this project, as was the ability of key staff members involved on the project to rapidly cross-train to develop skills in all necessary facets of IT (Eckerson, 2010) required to enable the Oracle warehouse to be conceptualized, modeled, built and maintained by UPS - with very limited IT support. This ownership of the warehouse by the key users and data specialists of the data, mean that it can respond very rapidly to additional requirements as there is no cross-functional, multi-level data warehouse governance structure is place (Watson et al, 2004).

The ability to incorporate the new HEPPP participation methodology into the warehouse is best done by staff who understand the intricate complexities of the individual components of the formula and how the data will need to presented for analysis. With a structured approach to development, this would be very time-consuming, as it is very difficult to provide detailed specifications to non-users of the data, when the exact methodology is still in the process of being determined (and provided by DEEWR) and there are still so many unknown factors.

Whilst there is a need for an enterprise-wide data warehouse at Monash to address the entire university's data needs, UPS had very specific information requirements which necessitated the development of the UPS data warehouse. The Monash University BI implementation is currently finalizing the second of its five portfolio area (Research, Education, HR, Finance and Advancement) implementations and will not be in a position to provide UPS with its very specific data requirements for a number of years.

The ever-present need to provide official, government-reported census data (upon which funding is based), with the ability to replicate published methodologies, means that UPS must have a centralized data repository which is able to rapidly respond to changes in the sector and grow and evolve like the reporting environment around it. With very limited resources UPS was able to develop a robust and integrated warehouse which was able to consolidate a wide range of data, significantly streamline the data preparation activities within UPS and enable the implementation of an integrated data management strategy within the department.

REFERENCES:

Australian Universities Quality Agency (AUQA) Audit Manual Version 7.1 (2010), Retrieved 20 September, 2010 from http://www.auqa.edu.au/files/auditmanuals/audit_manual_version_7.1_webversion.pdf

Behrangi, M & Fattolahi, A & Watson, H (2007) Determining Information Requirements for BI Applications. Business Intelligence Journal Vol. 12, No. 3, pp. 24-29.

Department of Education, Employment and Workplace Relations, (2009a) Mission-Based Compacts for Universities – A Framework for Discussion, Retrieved 20 September, 2010 from http://www.deewr.gov.au/HigherEducation/Policy/Documents/CompactsDiscussionPaper.pdf

Department of Education, Employment and Workplace Relations (2009b) Transforming Australia's Higher Education System, Retrieved 20 September, 2010 from http://www.deewr.gov.au/HigherEducation/Documents/PDF/Additional%20Report%20-%20Transforming%20Aus%20Higher%20ED_webaw.pdf

Department of Education, Employment and Workplace Relations (2010a) HEPPP Guidelines. Retrieved 20 September, 2010 from http://www.deewr.gov.au/HigherEducation/Programs/Equity/Documents/HEPPPGuidelines_2010.pdf

Department of Education, Employment and Workplace Relations (2010b) HEPPP initial 2010 allocation tables. Retreived 20 September, 2010 from http://www.deewr.gov.au/HigherEducation/Programs/Equity/Documents/HEPPPalloctableJun2010.pdf

Eckerson, W (2010) Revolutionary Business Intelligence: When Agile is Not Fast Enough. Retrieved 20 September 2010 from http://tdwi.org/articles/2010/04/08/experts-revolutionary-bi.aspx

Gibson, M & Arnott, D (2010) A participatory Case Study of Business Intelligence Systems Development, Bridging the Socio-technical Gap in Decision Support Systems – Challenges for the Next Decade, Volume 212, pp. 199-210.

Kakatani, K & Chuang, T (2005) The Development of a Data Mart at a Public Institution, Journal of Information Technology Case and Application Research, Vol. 7, No. 4, pp.30-52.

Kimball, R & Reeves, L & Ross, M & Thornthwaite, W (2002) The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses. Wiley Computer Publishing.

Monash University Business Intelligence Strategy document, (2006a). Retrieved from http://its.monash.edu/staff/projects/bi/documents/monash-university-bi-strategy-v1.1.pdf

Monash University Enhanced Data Dissemination Detailed Project Plan (2006b).

Monash University Enhanced Data Dissemination Project (EDDP) – Project Charter (2007).

Watson, H., Fuller, C., and Ariyachandra, T. (2004) Data Warehouse Governance: Best Practices at Blue Cross and Blue Shield of North Carolina, Decision Support Systems Vol. 38, No. 4, pp 435-450.

Wikipedia – Article on Dimension (data warehouse). Retrieved 20 September 2010 from http://en.wikipedia.org/wiki/Dimension_(data_warehouse)