

How Good is Your Data?

Authors: Andrea Matulick (The University of Adelaide),
Lachlan Murdoch (University of South Australia)
Presenters: Andrea Matulick (The University of Adelaide),
Lachlan Murdoch (University of South Australia)

Contents

1. Introduction
2. Data Quality Concepts
 - 2.1 Data as a Model
 - 2.2 Importance of Data Quality
 - 2.3 Defining and Assessing Data Quality
 - 2.4 Challenges in Improving Data Quality
 - 2.5 Establishing a Data Quality Framework
3. Data Quality at Universities
 - 3.1 Typical Data Usage Cycle for Universities
4. Examples of Data Quality Issues at Universities
 - 4.1 Online Student Enrolment Data
 - 4.2 Student Success Data
 - 4.3 Course Discipline Classification Data
5. The Effect of Change on Data Quality
6. Conclusion - Improving Data Quality
7. Abbreviations
8. References
9. About the Authors

1. Introduction

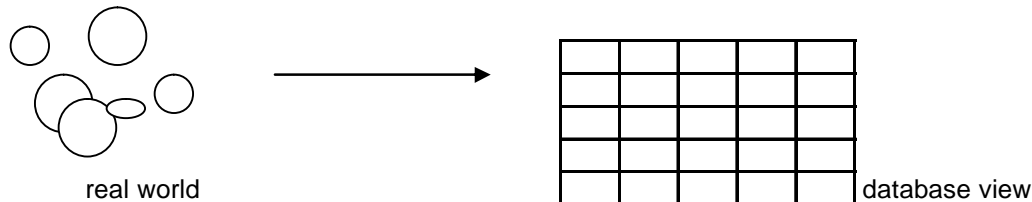
The arrival of the information age has resulted in massive amounts of data being stored within organisations, and this data becoming increasingly accessible through more sophisticated communication technologies and integrated information architecture. The recent growth of Data Warehousing demonstrates the degree to which organisations are now seeking to improve their return on their data by sharing information across applications and organisational units. However this centralisation of data also brings increasing focus on its quality. Universities too are managing increasingly large and complex amounts of data, not only to manage their business processes and comply with regulatory requirements, but also to gain increased value in order to manage more effectively and improve corporate performance. In this environment the focus on data quality has probably never been so intense. This paper examines concepts of data quality, the challenges and methods of improving data quality and data quality issues and examples in Universities.

2. General data quality concepts

2.1 Data as a Model

Data is an attempt to describe reality by capturing selected elements of the real world. Any such description is always partial and always only one of many possible “views” of reality.

Figure 1 Data as a Model of Reality



The unstructured, free-flowing form of the real world is mapped and categorised into our structured database model. Any given real world process can be modelled by many different data views, depending on one’s purpose. For example the speed of a ship might be represented by a single empirical value (e.g., 20 knots), by a relative value (e.g., slow, medium fast), by a table of numerical values representing speeds under different sea and load conditions, by the distance travelled relative to the total journey, etc. Similarly in Higher Education a student’s load might be represented by EFTSU this semester, this year or since commencement; as full-time or part-time study; by an academic monitoring view of which subjects have been completed and which are yet to be undertaken; by success rates; debts or fees incurred, etc. In each case a single aspect of reality (a student’s academic load) is represented by one or more data values. The choice of what kind of data to use to model reality is made prior to generating the data and the appropriateness of this choice plays a crucial role in determining the quality of the resulting data regardless of what data values are actually recorded (Rothenberg, 1996).

2.2 Importance of Data Quality

Information is a key organisational strategic asset, with critical business decisions and allocation of resources usually entirely based on data. However it is still often easier to demonstrate the costs or outcomes of poor quality data than the value of high quality data. For example American proprietary studies indicate that poor quality data costs 10% of organisations’ total revenue, and it has also been estimated the cost of poor quality data to US business exceed \$600 billion per year (Loshin, 2004). Also within the US, two pieces of legislation have been passed in the last few years that impose strict information quality guidelines on both public corporations (Sarbanes-Oxley Act of 2002) as well as US federal agencies (Data Quality Act of 2001). The Data Quality Act attempts to ensure that federal agencies use and disseminate accurate information, and reinforces the duty of Government to treat the information it collects and produces as a valued resource and that the information is secure, accurate and used appropriately.

2.3 Defining and Assessing Data Quality

There are two general concepts of data quality; the first is the “correctness” of the data, which most people think of in discussions about data quality, and include the concepts of

accuracy and consistency; and the second is the relevance of data for an intended purpose.

Data accuracy or correctness is essentially an intrinsic dimension of data quality, and typically is assessed independently of how the data is used. In the past, focus has often centred on this aspect of data quality without considering the context in which data is stored and used. However, increasingly it is being accepted that quality cannot be assessed independently of the people who use data – the data consumers. High quality data may therefore be defined as data that is fit for use by data consumers (Strong, 1997). English (1999) further clarifies this widely adopted definition by emphasizing that data quality is “fitness for all purposes in the enterprise processes that require it”, thus indicating that data quality could have a different meaning for different consumers. For example, data considered acceptable by one user may be unacceptable to another with more stringent requirements. In such a case the lack of acceptability may not be due to the “correctness” of the data but perhaps to its timeliness or level of detail.

Therefore data quality is a multi dimensional concept involving both the subjective perceptions of the individuals involved with the data, and the objective measurements based on the data set in question (Pipino et al, 2002). The subjective data quality assessments reflect the needs and experiences of stakeholders who collect, store or use data products – if they assess the quality of data as poor, their behaviour will be influenced by this assessment. Objective measurements reflect states of the data that may be task-dependent or task-independent. The range of dimensions that may be included in assessing data quality are outlined in Table 1.

Table 1

Dimensions	Definitions
Accessibility	data is available, or easily and quickly retrievable
Appropriate amount of data	volume of data is appropriate for the task at hand
Believability	data is regarded as true and credible
Completeness	data is not missing and is of sufficient depth and breadth for the task at hand
Concise representation	data is compactly represented
Consistent representation	data is presented in the same format
Ease of manipulation	data is easy to manipulate and apply to different tasks
Free of error	data is correct and reliable
Interpretability	data is in appropriate languages, symbols and units, and definitions are clear
Objectivity	data is unbiased, unprejudiced and impartial
Relevancy	data is applicable and helpful for the task at hand
Reputation	data is highly regarded in terms of its source and content
Security	access to data is restricted appropriately to maintain its security
Timeliness	data is sufficiently up to date for the task at hand
Understandability	data is easily comprehended
Value-added	data is beneficial and provides advantage from its use

Pipino et al, 2002

2.4 Challenges in Improving Data Quality

Central to the process of improving data quality is defining its importance to an organisation, especially given the difficulty in demonstrating a clear return-on-investment of improving data quality technology and processes. It is generally not clear what dollar amount can be placed on data that is inaccurate or missing from a dataset, or on data that is not available on a timely basis or in an appropriate form, in order to inform decision making. For example, what is the value of a report that is 99% accurate versus one that is 90% accurate, and how can you determine this measurement of accuracy? How can the extra effort and cost in improving data be quantified?

There is also the problem of data ownership; as discussed earlier, information quality is defined from the perspective of the consumer, yet the consumer does not control the generation of the information. Often there is a large disconnect between where the data is entered or captured and where it is used to make decisions. Additionally, although it is clearly preferable to find and correct problems at the source, in practice there may be reluctance or a certain lack of inertia in carrying this out effectively. Ultimately it is important to recognise that data quality is about people and processes rather than technology, although of course technology can facilitate improvement in these processes.

2.5 Establishing a Data Quality Framework

Information is a product that can be actively managed, measured and resourced like any other product of an organisation. Similarly a data quality framework can be established in the same way that other quality management processes are implemented, using similar methods of assessing the quality of the product (data), identifying data quality priorities and contributing to the continual improvement in data quality. Importantly, a data quality framework provides an enterprise approach to data quality management so that all areas within an organisation apply a common, objective approach to maintaining and improving data quality. An example of a benchmark data quality framework is given in the following case study.

Case Study: Canadian Institute for Health Information (CIHI)

Health services are increasingly reliant on data captured in clinical, administrative and management settings to manage service delivery, improve service performance and to develop policy (Paua Interface, 2004). Quality data in health care is especially important as life and death decisions may be involved.

In Canada a single organisation (CIHI) is responsible for maintaining health information. The Data Quality Framework was implemented in 2001 as a common strategy for assessing data quality across CIHI databases and registries. The core component of the framework is the Assessment Tool, which assesses and documents the limitations and strengths of a database. Once these issues have been identified they can be used to formulate recommendations for improvement, or strong points can be identified as best practice and shared between product areas.

Operationally the framework comprises 5 general dimensions, 19 characteristics, and 58 criteria - each dimension is made up of related characteristics and each characteristic is operationalised using the detailed criteria. These criteria are specific statements that, once assessed, determine how a particular characteristic is rated (CIHI, 2003). While the Assessment Tool provides standard definitions and a common strategy, the framework itself is designed to be part of a work process that identifies data quality priorities and produces continual improvement in data quality.

The five dimensions defined are:

Accuracy: how well information within a database reflects what was supposed to be reflected.

Comparability: extent to which a database can be properly integrated within the organisation's entire information system.

Timeliness: whether the data is available for user needs within a reasonable time period.

Usability: how easily the storage and documentation of data allows one to make intelligent use of the data.

Relevance: incorporating above elements to some degree, but focussing on value and adaptability.

The structure of the Assessment Tool listing the characteristics making up each dimension is shown in Table 2.

Table 2 The CIHI Data Quality Framework Evaluation Instrument

Dimension	Characteristics	No. Criteria	Dimension	Characteristics	No. Criteria
Accuracy	Coverage	4	Comparability	Data dictionary standards	2
	Capture & collection	5		Standardisation	2
	Unit non-response	3		Linkage	4
	Item non-reponse	2		Equivalency	2
	Measurement error	3		Historical comparability	3
	Edit & imputation	4	Usability	Accessibility	3
Estimation & processing	5	Documentation		3	
Timeliness	Data currency	4		currency	3
	Documentation			Interpretability	2
	currency	2	Relevance	Adaptability	2
				Value	3

Evaluation of the framework found it to be relatively strong theoretically as well as practical and reasonably generic, however several aspects were slated for improvement, such as more explanation of the trade offs involved across quality dimensions (CIHI, 2003).

3. Data Quality at Universities

Higher Education institutions are challenged with the task of providing large amounts of data to the government and other regulatory bodies. The data determines their funding and is used for comparison, benchmarking and ranking in various national publications. Internally, universities rely on this and many other sources of data as a management information tool to determine current performance, project future positions and check whether strategic directions and goals are being achieved.

To a large extent the data determines how each university 'looks' to the outside world including prospective students, researchers and employers. The truthfulness of the current picture of an organisation and accuracy of projections used for planning and development decisions is largely dependent on the quality of the data. Although collecting, aggregating and analysing the data to the stage of being useful for decision making is a time consuming task, time spent on checking the quality of the data is an equally important one which can be forgotten in the rush to get the output required. Additionally, the new Institution Assessment Framework (IAF) for ensuring accountability, quality and fairness

will be using quantitative and qualitative data from universities and external sources to assess institutions. This will place more pressure on universities to be aware of the quality of their data.

As stored data can only ever be a representation of the real world, problems occur in data entry where real data does not fit exactly (or at all) into the boxes created to represent it. Compromises need to be made so approximate data can be entered. Some real data can be totally different to the rules created to check data consistency. (For example, persons of ATSI descent who were not born in Australia.) A decision must be made whether to accept the real data and allow the rule to be broken or modify the data slightly to allow the data to follow the rule. If instances of a particular inconsistency increase, rules may need to be modified over time.

3.1 Typical Data Usage Cycle for Universities

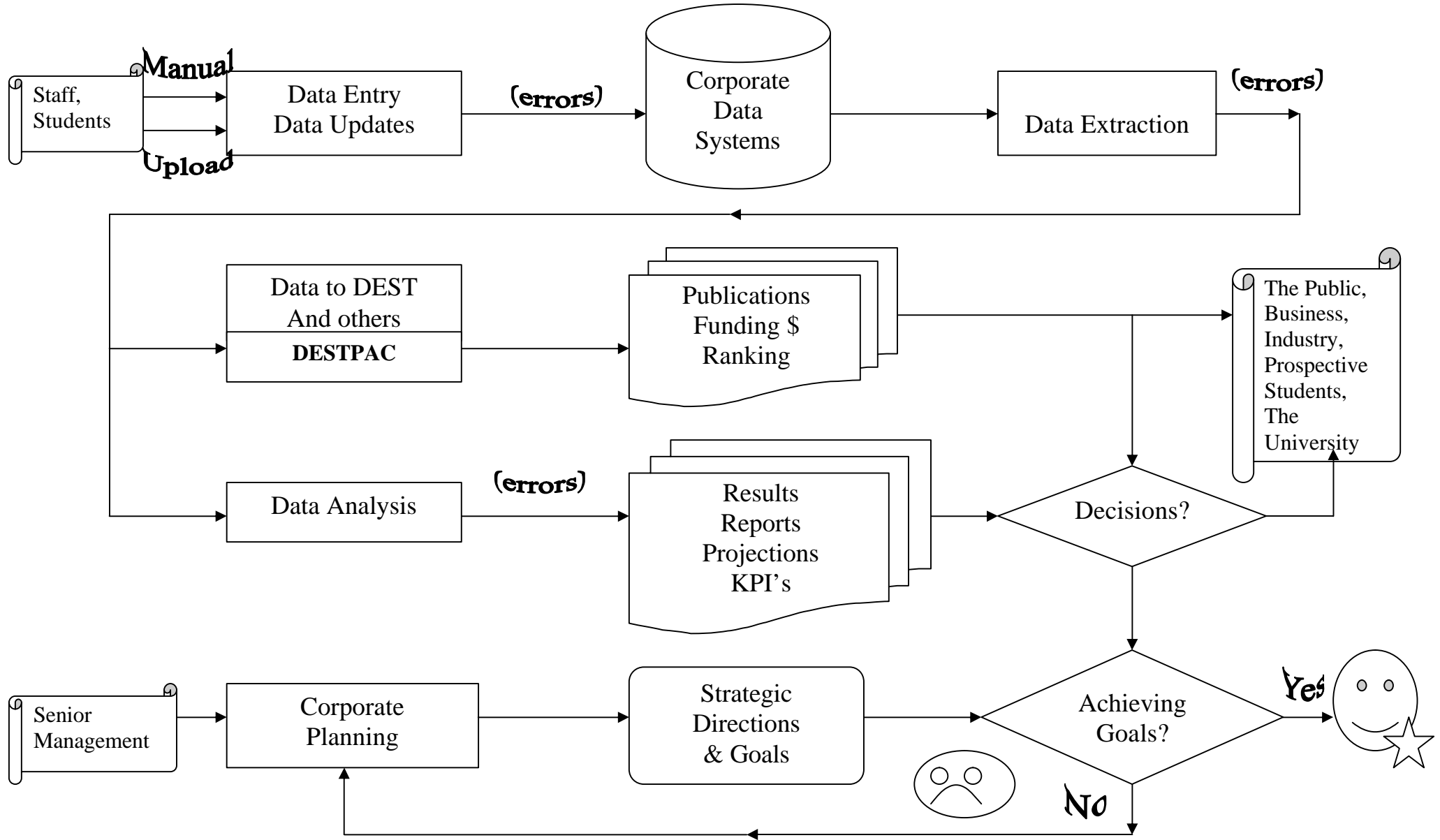
Figure 2 shows a typical data cycle within a university. It shows the sources of data, how the data is stored, how it is extracted and analysed to produce reports, results and projections for the university, which in turn are used by management for decision making, and by regulatory bodies to determine performance, funding and rankings. The results of analysis are compared with strategic goals and directions of the institution. Key corporate decisions can be driven by the results. If errors in the data produce misleading results, decisions can be costly by way of economic damage, lost opportunities, bad publicity and risk to reputation. Therefore data should be treated as a strategic corporate resource (Bushell, 2003).

Typical data for universities include student and staff bio-demographic data and life-cycle data during their presence at the university (e.g. student enrolment and results data, staff positions, classification and contract type), financial transactions data (income from government funding, student fees, research grants, expenditure for staff salaries) and survey data (Graduate Destination Survey (GDS), Course Experience Questionnaire (CEQ), Student Evaluation of Learning and Teaching (SELT), etc.)

The data is stored in various systems or locations around the university. Most institutions have a student record administration system, human resources (HR) record system and finance system. Many have a research management system. Other stand-alone systems may exist to store data about applications, international students, scholarships, cooperative research entities and evaluation activities. Increasingly, this diverse set of data is being integrated into a data warehouse or datamarts that permit cross functional reporting.

Data is entered into the various systems either manually or by uploading data from another system. Manual data entry can be handled in batch form by data entry staff at the institution (usually from paper forms), or individually by students and/or staff using online facilities via a website. Data uploads may be obtained from other internal or external computer systems. A large amount of application data is usually received from a Tertiary Admission Centre (TAC). A Research Master database may get a personnel update regularly from the HR system. The method of data entry can dictate the level of data quality achieved. Generally, manual entry is less accurate than uploads of data from another system, but only if the uploaded data has been validated at some stage. Studies have shown that errors for manually entered data entry result in about 5% inaccuracies and 5% missing data (ARTS et al, 2002).

FIGURE 2: Typical Data Cycle for Universities



Data is extracted from the various systems for sending to regulatory bodies as statutory reports and also for internal analysis into standard management information reports, or to produce key performance indicators. The more difficult analysis often involves data from more than one system needing to be integrated to produce the required report. (For example research income by staff member, or cost analysis of teaching a course). Errors can be introduced into the data at the time of extraction due to a number of reasons. The data required for analysis may not be stored in the system in the form required (For example in Peoplesoft systems, commence date is not stored for undergraduate students). Some form of assumption or approximation may be required. An incorrect filtering or extraction algorithm may be used. All data may not be included in the extraction. (For example international incoming exchange students are not included in the DEST student data, and many offshore teaching staff are not included in staff data). Inaccuracies introduced at the time of extraction are thought to be about 2%(ARTS et al, 2002)

Providing clear documentation which maps the processes producing data is very important. DEST provides comprehensive documentation and rigorous data specifications for its student and staff data extractions. It provides a software application called DESTPAC which carries out numerous validation checks on the data and can automatically produce many reports including variation reports compared with previous years. Although not perfect, a number of universities have relied on DESTPAC to form an important part of their data validation. One disadvantage of DESTPAC is that validating data 'on output' from the university system is not as efficient as validating 'on input' as will be shown later. Some universities regularly and automatically generate commonly used reports. Others process their raw data into 'value added' products such as web enabled data warehouses and pivot tables which provide users with the ability to produce their own reports.

Regularly used reports can be valuable in indicating areas of poor data quality. They are likely to be more accurate than an ad-hoc analysis on raw data. Comparison and variation reports which produce more than a 10% variation over previous data should be investigated and explained. Trend analyses, which produce unexpected sudden increases or decreases, may also indicate poor quality data rather than an impossible to maintain trend or decline. More errors are likely to creep in at the analysis stage for infrequent or one-of reports where the scope, data specifications and assumptions are not well defined and tested (Bushell, 2003).

When data quality problems are identified, they need to be prioritised as to the effect and cost to the organisation of the bad data. A few errors in frequently used critical data can be more costly than many errors in data that is rarely, if ever used. (For example a 1% error in load reporting to DEST could cost a university \$1million or more, whereas if 10% of staff highest qualification is missing, it may not be noticed or have any cost implications.)

4. Examples of Data Quality Issues at Universities

4.1 Online Student Enrolment Data

In the interests of increasing enrolment efficiency at universities, and to provide a better service to students, more and more universities are providing online enrolment facilities. For most universities, this involves a huge cultural and technological change from the

previous paper based enrolment method. Instead of thousands of students filling out their details on paper forms which are then checked (with varying degrees of rigor) by university staff, before being handed on for batch entry into the university system by yet more university staff, each student can access the online enrolment facility via the university's web site and enrol themselves.

In the past few years I have been involved at two universities which have implemented online enrolment and have experienced interesting reservations, expectations and outcomes which have been similar at each. For this paper I will concentrate on the effect on data quality as a result of the implementation. From my role as DEST Data Collection Coordinator at one university, I was aware of what I thought was an unnecessarily high amount of missing and incorrect data in the student statistical data collected for DEST. This was highlighted by the number of errors being produced by the DESTPAC validation tool provided by DEST. At this time, DEST student data was collected as a snapshot twice a year. The data was due at DEST two months after each census date, but most universities could start validating the data at least one month before the census date. This meant a university had about 12 weeks to validate the data.

On the surface, this seemed to me like ample time to make the necessary corrections to even vast amounts of data. So I was bemused to find that many universities struggled to finish the data on time and many went over the submission due dates. For a university of about 25,000 students, we had around 15,000 errors when we started validating our data in early March. After 12 weeks, we would usually still have about 500 errors to remove (a desperate last week). Table 3 shows the number of errors categorised by area and total errors over one validation period in 1999. I have also given some idea of the percentage of data records with errors. You can see there are over 3,000 errors in the Campus/Division area, many of which are errors in the student statistical data. Another area producing large numbers of errors is the Finance area, but these errors are not part of the discussion in this paper.

Table 3

**Summary of DETYA Errors - Submission 1, 1999
Before Online Enrolment - University with 25,000 students.**

Date	%Data Records with Errors	Total Errors	Registry/ Finance/ Planning Errors	Total Campus/ Division Errors	Missing Citizen- Resident Code	Missing HECS Code	
8-Mar	7.1%	14,236			815	506	
15-Mar	8.8%	17,664	14,438	3,226	800	499	
22-Mar	2.8%	5,612	2,278	3,334	687	397	
30-Mar	2.0%	4,055	1,398	2,657	438	245	Census Date 31/3
6-Apr	3.5%	6,942	4,988	1,954	351	154	
19-Apr	2.8%	5,513	4,145	1,368	335	140	
27-Apr	2.7%	5,310	4,104	1,206	318	110	
3-May	2.2%	4,422	3,570	852	223	98	
10-May	0.6%	1,245	548	697	161	56	
18-May	0.5%	902	na	na	190	86	
24-May	0.2%	496	244	252	55	Cancelled	Submission Date 28/5
31-May	0.1%	248	83	165	12		
7-Jun	0.0%	98	20	78	7		

I became aware that some of the DEST statistical data being requested was not considered particularly important by university staff as 'necessary' to enrol a student. This included things like citizenship, home and term address postcodes, year of arrival in Australia, prior education statistics and admission basis. In fact a few areas seemed to be happy with just a name to enrol their students. I was surprised to find that although the in-house student record system was fairly sophisticated, it had very few validation checks on data entry by university staff, and many data fields could be left blank. At first I quietly blamed the students for being lazy and not supplying the data, but as time went on I discovered that a lot of data was on the enrolment forms, but when staff were under pressure to get the forms entered on time, they knew they did not have to enter the statistical data, so they left it 'to do later'. Although previous data coordinators had 'fixed' a lot of errors themselves to get the submissions in on time, the current practice was now to return the data errors to the source for correction.

This was a very important step in making data quality everyone's responsibility, but incredibly staff would still wait until the missing data came up in errors and was sent back to them before looking up the paper enrolment forms to get the data. They did not seem to worry about the time delay and inefficiency this created, plus the frustration of getting long lists of errors that meant looking up large numbers of students again (and again). It seemed to me that adding code to our in-house system that would check the data at the time of entry and make it more difficult for staff to leave out statutory information would improve efficiency for many staff. I approached our IT system staff about this but was met with a lukewarm response. What if the checks stopped the staff from enrolling a student? How would they get the students enrolled on time? They were sure staff always entered the data if it was on the form, so the only missing data would be because the student had not filled it in. I pointed out that the number of students who did not appear to know what their citizenship, date of birth, and address was seemed alarmingly high, but this did not sway them either. I was also put fairly low down the priority list because the system team was busy working on implementing online enrolment....

Online enrolment was discussed at many of the data validation meetings held regularly about the DEST submission data. Faculty, school and enrolment office staff were very concerned about the amount of 'rubbish' data that would be entered by students if they were allowed to enrol themselves. They were sure the students would not know what courses to enrol in, and would enter silly things while trying to be funny. We would get hundreds of students called 'Sir' and 'Esq' and all the student addresses would be things like 'Parliament House, Canberra'. I started to become uneasy myself about how the new system would make it even harder to get our data submitted on time. I knew implementing a new system would have inevitable 'teething' problems, but even worse data quality in addition would be a nightmare. Another problem I could foresee was how were we going to find missing data from online enrolment? I was sure the staff would now wipe their hands of missing and incorrect stats data, as it was solely the students' responsibility. Would I be needing to contact hundreds of students personally? What if the students had all given incorrect contact details? Was it a good time to start looking for a different job?

As a member of the Project Implementation Reference group, I approached the system implementation consultants with a request to include data edit on entry as part of the online enrolment screens. They agreed in principle with the idea and I supplied some basic specifications about which data fields needed to be mandatory and which needed to be cross checked with other data to ensure consistent results. It was interesting that while it was deemed unacceptable for staff to be prevented from enrolling a student when data

was missing, there was no objection to stopping a student from enrolling if they did not enter the required data. However, we tried to specify the checks only to do this when it was very unlikely the student would not know the required data. Further cross checking for consistent data would be still carried out during the enrolment and the student would be asked to 'sign off' on the data as true and correct. Any TAC data already available in the system was populated into the enrolment screen. The student could update it if it was out of date or incorrect. This feature of online enrolment gave added convenience as it meant we were not asking the student to supply the same information twice.

As go-live day came and went I waited nervously to hear news about the success or otherwise of our first online student enrolments. The news was good. Online enrolments were a success, students were generally satisfied with the process and students having difficulties were being handled by the various methods put in place (Watson, 2004). The validation of student data for our next submission commenced. What affect had online enrolment had on our data quality? Our first extraction of data and validation by DESTPAC was quite amusing. We came out with over 1 million errors in our data for 25,000 students! Even students cannot make this many mistakes. The problem was in the vendor supplied extraction programs which had incorrect data structures for most of the files. Almost every data field had an error. Once this was rectified, we could see a better picture of our data. The most noticeable difference was the large drop in enrolment file errors. We had gone from thousands of errors to hundreds, a tenfold drop. This was a very pleasant surprise, and even although other teething problems kept us busy right up until the submission date, it showed that the work done in adding data editing on input to the online enrolment screens was a big success. Table 4 shows enrolment file (Enfile) error numbers before and after implementation of online enrolment for a university with about 17,000 students.

Table 4

**Summary of DEST Errors - Submission 1, March 2003 and 2004
Before and After Online Enrolment - University with 17,000 students.**

Date	%Data Records with Errors	Aofile	COfile	ENfile	LDile	LSfile	Total Errors
Before Online Enrolment							
19/03/2003	1.6%	0	38	1760	115	429	2342
30/05/2003	0.1%	0	0	113	32	25	170
After Online Enrolment							
19/03/2004	4.5%	6	1	331	5	6405	6748
28/05/2004	0.1%	0	0	126	0	35	161

An estimate of the amount of staff hours saved due to reduced error correction after the introduction of online enrolment would be about 300 staff hours per submission. (From 420 hours to 120 hours. 5 staff (1 per Faculty) spending 7 hours (1 day per week) for 12 weeks down to 5 staff spending 2 hours for 12 weeks). This is in addition to the huge amount of time saved because staff no longer enter the data from paper based forms. Another bonus noticed from the introduction of online enrolment was that staff had more time to spend on students who had genuine enrolment problems. Staff took more interest in the smaller

number of errors that did need to be followed up post-enrolment. Student contact details had also been improved due to the introduction of online address and phone number updating at about the same time. Giving students ownership of their data and staff more time to concentrate on those who needed help most gave everyone a greater sense of satisfaction. The expected problems with students entering 'rubbish' data had not eventuated. However, from the table you can see that student financial errors (Lsfile) were still a problem.

Conversion to student centred online enrolment, including data validation on input, produced immediate measurable and cost effective improvement in data quality. Giving students responsibility for their own data and staff the responsibility for only genuine 'problem' students had focussed the effort into a far more productive process.

4.2 Student Success Data

Student success is likely to become more important as an indicator of performance under the new Institution Assessment Framework (IAF) to commence in 2005 (DEST, 2004). At present, success is calculated as one of the Equity performance indicators for domestic students only. Tables are produced at university, state and national levels for benchmarking performance. DEST defines student progress rate as the student load passed as a proportion of the student load certified (attempted). So a student who sits for 4 subjects of equal credit and passes 3 would have a success ratio of 0.75. Subjects which are ongoing (e.g. research or perhaps honours) or for which no result is known are not included in the calculation. Subjects where the student withdraws without penalty are included in the load attempted but not passed. The passed and attempted load are reported in the 3rd submission of the student load file. This file is submitted in March for all load reported in the two submissions of the previous year. So data for 2003 load was reported to DEST in March 2004. Records are reported for each unit of study (subject) for each student. The 'unit of study completion status' is the data element used to report this data. This should be populated with final values by the March (following year) submission, based on the grade the student receives for each subject. The data element can take on one of 4 values as below:

- 1 Withdrew without penalty
- 2 Failed
- 3 Passed
- 4 Incomplete or unknown status

Student Progress Rate = load for 3 / load for (1+2+3)

As part of an exercise to produce reports on student success along with grade distribution for internal university use, our Planning Office arranged to have student mark and grade data included in the university filler of the 3rd Submission load file. Before analysing the data, it was decided to check the consistency of the unit of study completion status against the grades data added to the file. Table 5 shows the results of this comparison.

Grade data was extracted from the current and legacy student record systems for the previous 8 years. The data consisted of over 100,000 load records for each year with the 4 types of completion status, a mark between 0 and 100 and one of 29 different grades. (The usual pass, credit, distinction plus others such as conceded pass, incomplete fail, etc.) The grades were grouped into those that would match each of the 4 possible unit of

study completion status values. The number of grades that matched the completion status, and number that did not were grouped into 'consistent' and 'inconsistent'. The percentage of each group that was inconsistent was calculated. The 2002 data showed a very poor consistency. This was the first time the completion status data had been extracted from a newly implemented Peoplesoft student system. The cause was investigated and found to be the way grades for semesterised units of study had been extracted from the system. This was rectified for the 2003 data, producing an almost perfect data consistency. The error may have gone unnoticed if a time series had not been produced.

The largest inconsistency for most years was the data with no grade (4) at around 15%. This was most likely because the grade data had been extracted from the student systems some years after the completion status data had been submitted to DEST. Most of the inconsistency would come from missing grades being entered at a later date. The most consistent data was the success grades (3), usually less than 2% inconsistent. This was encouraging as it was by far the largest amount of data, and the main component in calculating the success ratio. An overall inconsistency of data for each year was found to be between 2 and 4%.

The results of the comparison indicated that data consistency was not 100%, and would not be so for the historic data. However, the data accuracy of each set was of high enough standard to accept both sets of data and present reports based on grades and success. It was decided not to present the data combined in one table for analysis by university staff because of the inconsistencies. Separate tables would be produced for success data and for grade distribution data. The exercise showed that analysis of data, which is supposedly from the same source, is another good way of checking the standard of data quality.

Table 5 Grade and Completion Status Distribution for DEST Submission 3 Load files 1995 - 2003

Year	UoS Compl Status	Success	No. of Grades	Grade consistent with Compl Status?		
				Consistent	Inconsistent	%Inconsistent
1995	1 (Withdrawn)		5136	4853	283	5.51%
	2 (Fail)		10876	10479	397	3.65%
	3 (Success)		85167	84667	500	0.59%
	4 (No Grade)		8481	7992	489	5.77%
1995 Total		0.84	109660	107991	1669	1.52%
1996	1 (Withdrawn)		3189	3013	176	5.52%
	2 (Fail)		13691	12012	1679	12.26%
	3 (Success)		87658	86873	785	0.90%
	4 (No Grade)		5248	3825	1423	27.12%
1996 Total		0.83	109786	105723	4063	3.70%
1997	1 (Withdrawn)		3874	3688	186	4.80%
	2 (Fail)		12374	11545	829	6.70%
	3 (Success)		90061	89294	767	0.85%
	4 (No Grade)		4549	3741	808	17.76%
1997 Total		0.85	110858	108268	2590	2.34%
1998	1 (Withdrawn)		3400	3183	217	6.38%
	2 (Fail)		12072	11410	662	5.48%
	3 (Success)		89947	88961	986	1.10%
	4 (No Grade)		4362	3743	619	14.19%
1998 Total		0.85	109781	107297	2484	2.26%
1999	1 (Withdrawn)		2981	2734	247	8.29%
	2 (Fail)		11697	11100	597	5.10%
	3 (Success)		88311	87292	1019	1.15%
	4 (No Grade)		4326	3680	646	14.93%
1999 Total		0.86	107315	104806	2509	2.34%
2000	1 (Withdrawn)		2452	2281	171	6.97%
	2 (Fail)		11250	10455	795	7.07%
	3 (Success)		82869	81966	903	1.09%
	4 (No Grade)		5076	4237	839	16.53%
2000 Total		0.86	101647	98939	2708	2.66%
2001	1 (Withdrawn)		2712	2037	675	24.89%
	2 (Fail)		11601	11427	174	1.50%
	3 (Success)		85999	84057	1942	2.26%
	4 (No Grade)		4750	3808	942	19.83%
2001 Total		0.84	105062	101329	3733	3.55%
2002	1 (Withdrawn)		2424	2410	14	0.58%
	2 (Fail)		12483	10634	1849	14.81%
	3 (Success)		89873	75598	14275	15.88%
	4 (No Grade)		4691	4056	635	13.54%
2002 Total		0.86	109471	92698	16773	15.32%
2003 Orig	1 (Withdrawn)		2024	2013	11	0.54%
	2 (Fail)		12344	10901	1443	11.69%
	3 (Success)		93791	81292	12499	13.33%
	4 (No Grade)		5433	5432	1	0.02%
2003 Total Orig			113592	99638	13954	12.28%
2003 Fixed	1 (Withdrawn)		2034	2034	0	0.00%
	2 (Fail)		12307	12306	1	0.01%
	3 (Success)		94082	94021	61	0.06%
	4 (No Grade)		5282	5281	1	0.02%
2003 Total Fixed		0.87	113705	113642	63	0.06%

4.3 Course Discipline Classification Data

Under the new Commonwealth Grant Scheme (CGS) from 2005, universities will be far more closely monitored to ensure they meet their commonwealth funded student load as per the discipline cluster mix in their funding agreement. Universities will still be fully funded up to 1% over their total target load at the agreed mix, and may carry load up to 5% over the target before receiving penalties for being overloaded. Universities who go under the target for more than one year are likely to have their load target and funding reduced. The funding agreement is based on the 12 clusters below.

CGS Cluster	CGS Discipline
01	Law
02	Accounting, Administration, Economics, Commerce
03	Humanities
04	Mathematics, Statistics
05	Behavioural Science, Social Studies
06	Computing, Built Environment, Health
07	Foreign Languages, Visual and Performing Arts
08	Engineering, Science, Surveying
09	Dentistry, Medicine, Veterinary Science
10	Agriculture
11	Education (national priority)
12	Nursing (national priority)

The load in each cluster is determined by the discipline code reported against each student load record for each student in each unit of study (i.e. subject or course). The discipline code is a Field of Education (FOE) code as defined by the Australian Bureau of Statistics – Australian Standard Classification of Education (ASCED). This same discipline code determines the HECS band level at which a student is charged their student contribution amount. The discipline should be allocated according to the academic content of the unit of study, not necessarily the same as the main discipline of the area that teaches the unit. Discipline has been reported against unit of study since 1997 to determine which of the 3 differential HECS bands applies. In some universities it is also used to weight teaching load for internal budget distribution of commonwealth funds to schools or departments. However, for other institutions it has not been critical to get every subject classified to the correct discipline within one of 12 cluster groups.

The introduction of funding agreements with such small tolerances on load within discipline cluster prompted my university to carry out a discipline review for all courses for 2004. This was a large exercise as over 4,000 courses were listed at the time the review was commenced. Lists of courses, sorted by the school with administrative responsibility for the course and the current discipline (FOE) classification, were sent to each Faculty. The number of different FOE's that can be allocated is about 440 but these are grouped into 12 broad Fields of Education, similar to (although not the same) as the 12 funding cluster groups, except for the national priorities of teaching and nursing. The ASCED documentation, and a link to an electronic version suitable for searching, was sent to each Faculty as the documentation explains in more detail with examples the sort of subject content that should be classified to each field of education.

Faculties were given about 6 weeks to review the classification of their courses, and while most already had an understanding about the soon-to-be-stronger link between the classification and the funding, we were pleased to find that all Faculties went about the task in a very professional manner and were quite fastidious to see that their courses were classified correctly according to ASCED. We also tried to mention the words 'possible audit' occasionally in general discussions about reported data.

To test the results of the review, course classification before and after the review was applied to our full year 2003 load data. This included load in 1925 courses. Of these, 297 had changed discipline (FOE) after the review (15%). Of these, 101 resulted in changed CGS Cluster (5%). The load in these 101 courses accounted for 7% of our total 2003 load. 25 of the courses accounted for 80% of the load that changed cluster.

The main changes required to course classifications were:

- Statistics, Maths and Information Technology courses taught by the Schools of Commerce and Economics had been classified to business disciplines.
- Commercial and Taxation Law courses had been classified to Business Management.
- Audio Visual Studies courses had been classified to general Humanities disciplines.
- Environmental Studies courses had been classified to Human Geography or Studies in Human Society.
- Education Psychology courses had been classified to Education, whereas they are specifically excluded in the ASCED documentation.

The change in load for each cluster as a percentage of the total load ranged from a decrease of 3.3% in Cluster 02 (Accounting, Administration, Economics, Commerce) to an increase of 1.5% in Cluster 07 (Foreign Languages, Visual and Performing Arts). Cluster 04 (Mathematics, Statistics) increased by 1.4% and Cluster 01 (Law) by 1.2%. All other changes were less than 1%. Table 6 and Figure 2 show the results of the cluster redistribution of load.

Figure 3 shows the change in 'value' of the Commonwealth Contribution funding when applied to the two distributions. (The student contribution has not been included.) The new 'value' of the same load is \$1.8 million more than before the cluster review was carried out. This difference comprised increases totaling almost \$4.0 million over 7 of the clusters, and decreases of \$2.1 million in another 4 clusters. \$1.8 million represents 2.5% of the 'value' of the load before the review was carried out. Unfortunately, as DEST's base funding agreement for each university is based on their 2003 reported load (before the new Act was passed and the review was carried out), the reclassification does not mean the university gains any more funding from the exercise. The university was substantially overloaded in 2003, so we were not being funded for the full amount of load even at the 2003 base. However, the review did mean we were able to re-negotiate our cluster distribution more closely to the reclassified load, which will help us to meet load targets as we move into the future. (The re-negotiation was possible because we redistributed the load to increase our total target by a number of places while still retaining the same commonwealth funding target of load.)

The rightmost column in Table 6 shows the change in commonwealth funding value of each cluster as a percentage of the value before the review. Although the total change in funding value was only 2.5%, changes in cluster funding varied from +29% in Cluster 04 (Mathematics, Statistics) to -20% in Cluster 02 (Accounting, Administration, Economics, Commerce). If the CGS cluster distribution is used for internal funding allocation, these substantial changes would need to be addressed in some way.

Table 6 Results of Discipline and Cluster Review, using 2003 EFTSU

Cluster Description	EFTSU Before review	EFTSU After Review	EFTSU Change	EFTSU %Change	%Total Before Review	%Total After Review	Total %Change	Comm Cont (+2.5%) Before	Comm Cont (+2.5%) After	Comm Cont Change	%Change for Cluster
01 Law	802.6	917.2	114.6	14.3%	8.1%	9.3%	1.2%	\$1,210,952	\$1,383,837	\$172,885	14.3%
02 Acct, Admin, Econ, Comm	1600.0	1274.6	-325.4	-20.3%	16.1%	12.9%	-3.3%	\$3,968,791	\$3,161,665	-\$807,126	-20.3%
03 Humanities	883.7	800.0	-83.7	-9.5%	8.9%	8.1%	-0.8%	\$3,693,704	\$3,343,978	-\$349,726	-9.5%
04 Maths, Stats	475.6	615.1	139.5	29.3%	4.8%	6.2%	1.4%	\$2,348,022	\$3,037,002	\$688,980	29.3%
05 Behav Sci, Social Stud	1400.2	1291.0	-109.2	-7.8%	14.1%	13.0%	-1.1%	\$9,293,251	\$8,568,243	-\$725,009	-7.8%
06 Comp, Built Env, Health	589.8	660.3	70.5	12.0%	5.9%	6.7%	0.7%	\$4,359,890	\$4,881,358	\$521,468	12.0%
07 Foreign Lang, Vis and Perf Arts	743.0	889.0	146.0	19.7%	7.5%	9.0%	1.5%	\$6,754,347	\$8,081,982	\$1,327,635	19.7%
08 Eng, Sci, Surv	2221.6	2229.9	8.3	0.4%	22.4%	22.5%	0.1%	\$27,332,394	\$27,434,895	\$102,502	0.4%
09 Dent, Med, Vet Sci	726.5	729.4	2.9	0.4%	7.3%	7.4%	0.0%	\$11,205,241	\$11,249,614	\$44,374	0.4%
10 Agriculture	292.5	361.9	69.4	23.7%	3.0%	3.7%	0.7%	\$4,796,147	\$5,933,620	\$1,137,473	23.7%
11 Education	177.4	144.4	-33.0	-18.6%	1.8%	1.5%	-0.3%	\$1,294,241	\$1,053,542	-\$240,699	-18.6%
12 Nursing	1.5	1.5	0.0	0.0%	0.0%	0.0%	0.0%	\$14,623	\$14,623	\$0	0.0%
Total	9914.4	9914.4	0.0	0.0%	100.0%	100.0%	0.0%	\$76,271,603	\$78,144,359	\$1,872,756	2.5%

Figure 2 Discipline and Cluster Review, %Change in Total EFTSU by CGS Cluster

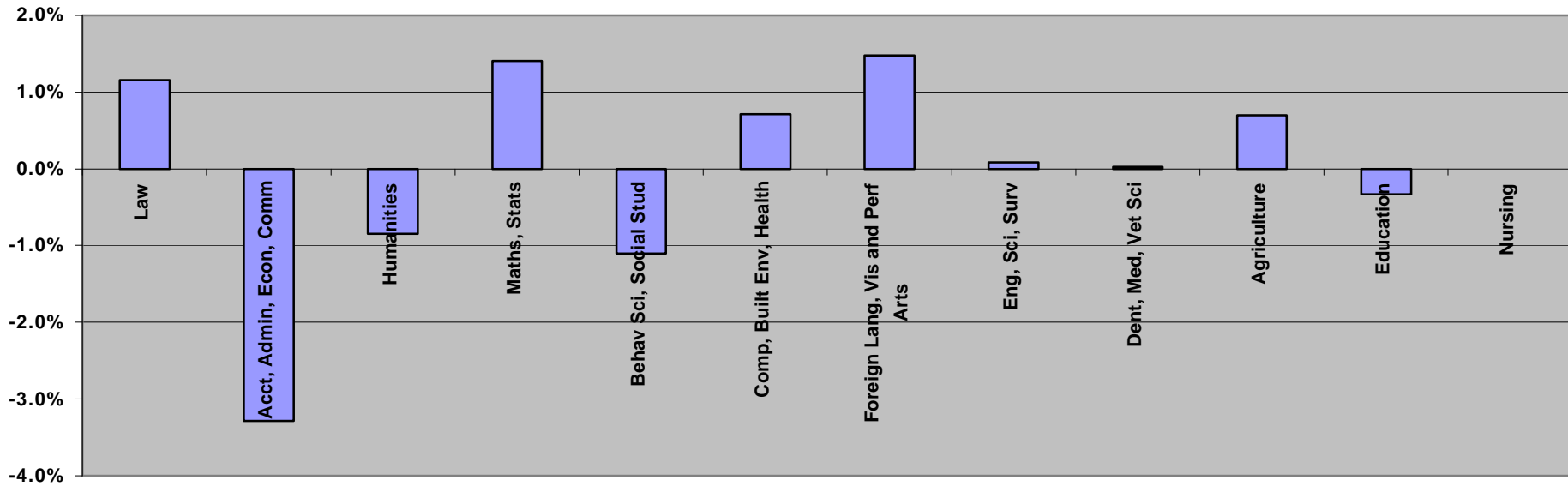
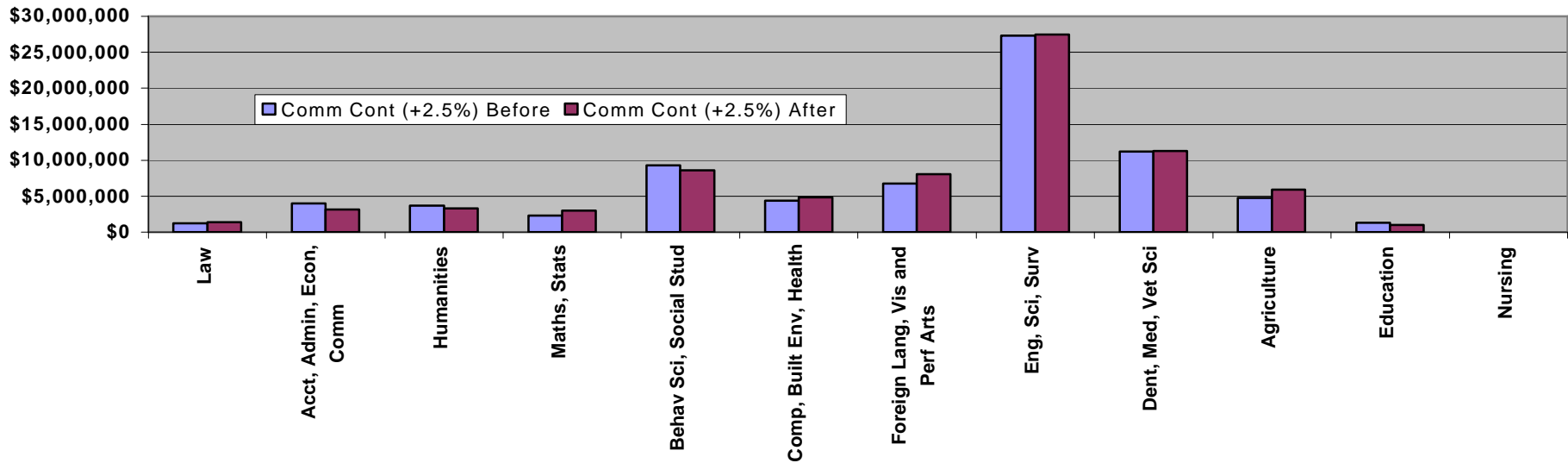


Figure 3. Discipline and Cluster Review, Change in Comm Funding 'value'



The result of the review of discipline classification of courses showed that although about 95% of courses are easy to classify, the last 5% can be more difficult, with a few very difficult to classify. Incorrect classifications in heavily enrolled courses that result in a change of CGS cluster can make a substantial difference to the 'value' of the load. The data quality issue here is about trying to classify real world information into a finite set of specified descriptions. Now that discipline classification is so closely tied to funding, higher education providers will need to have processes in place to assure the quality of this data.

5. The Effect of Change on Data Quality

Introducing changes to a system or process is usually accompanied by initial data quality problems.

Systems and business processes that have been stable for a reasonable period are likely to have a higher level of data quality as minor corrections and improvements have been made over time. However, some data quality problems that have been labelled 'too hard' may become endemic in a system and require a large input of resources (including financial) to correct. (e.g. the general ledger in a financial system, or changing student billing type from semester/term level to unit of study level.) In these cases, the cost of low quality data must be carefully compared with the cost to 'rectify' the problem before a decision is made.

Using centralised data systems wherever possible and only having one data source for each type of data also helps improve data quality. For example, limit stand alone systems and where they have to exist, use regular uploads from major systems for any data that is duplicated in both systems (e.g. student and staff ID's).

Changes in data specifications can cause temporary fluctuations in data quality. For instance, DEST is about to radically change the scope and structure of student data to be reported. An equivalent of about ten files will be required instead of the current five and the timing of submissions is increasing from the current 2 or 3 submissions to eventually being continuous. Although a few data elements have been removed, at least 30 new elements are required. A large number of these new elements have only just been defined (as a result of the new Act) and many are not currently stored by university systems. With this scale of changes, institutions will probably need to allow 1-2 years to bed down new processes and iron out data quality issues.

Implementing a new corporate information system at an institution is also a time to carefully monitor data quality. Changes in methods of data entry and new business processes can be detrimental to data quality as staff, students and others go through a 'learning curve'. Adequate time for testing, correcting and fine tuning should always be included as part of the implementation schedule for changes in business processes, specifications or computerised systems.

6. Conclusion - Improving Data Quality

We have seen by the above discussion and examples that data quality can be actively managed. Higher Education institutions should not be daunted by the task of monitoring and improving the quality of the large amounts of data they collect. A framework of processes can be set up to manage data quality. All staff must be included in high level buy in and some given clear data governance responsibilities.

Staff and students should be given ownership and responsibility for their data. They should receive feedback about problems and improvements as they occur or are identified. Data issues and possible solutions should be prioritised, taking into account the size of the problem and the cost of implementing a solution (both outlay and savings).

The main areas where errors are introduced into data are at entry or upload, during extraction for reports or during analysis. Data validation on input is an extremely effective method of improving data quality. This has been evident at universities who have implemented it as part of an online enrolment system. Data validation on output (e.g. DESTPAC) is also a useful tool, but not as efficient. Having only one source for each data item, and using finite selective lists (drop down lists) for text and code fields where possible helps to produce more accurate and consistent data.

Developing and using standard definitions and processes which are consistent across business areas entrenches better data quality. These definitions and processes should be well documented and updated regularly. Using regular automatic extracts to produce reports, and monitoring over time is a good way to uncover problems with data quality or incorrect extraction routines. Significant improvements in consistency can be achieved, as shown with the example of completion status vs student grades.

Classification of data to a standard set of codes can produce difficulties, as with the discipline group and cluster classification of courses. Staff should be supplied with the full documentation that applies to the classification and given assistance with the small percent of cases which are difficult. Where these classifications are linked to funding, small improvements can have significant value implications.

Be aware that changes can affect data quality over the short or long term. Changing business processes, implementing a new system, changing data definitions or extractions and timing of reports are instances where data quality may deteriorate. Always allow adequate time for testing and checking during times of change.

Make your data visible and usable by producing publicly accessible pivot tables or online reports, web enabled reporting or a data warehouse. Take note of usage reports and web page 'hits'. Provide a feedback form and contact details for queries about the data. Knowing that many people can see and do use your data will give you confidence and even pride in it.

Using a combination of all the above to produce a sound data quality framework based on people and processes which is continually monitored will ensure improvement and quality results.

Ten Tips for Improving Data Quality

1. Establish a framework for managing data quality
2. Include staff in high level buy in with clear data governance responsibilities
3. Develop or use standard definitions and processes which are consistent across business units
4. Give staff (and students) ownership and responsibility for relevant data, return corrections to the source
5. Continuously acknowledge improvements and provide feedback to problem areas
6. Prioritise the data issues and the solutions (including relative costs)
7. Validate data on input (garbage in, garbage out)
8. Have only one source for each data item (avoid free text and use drop down lists where possible)
9. Use regular extractions and reports and monitor changes over time
10. Make your data visible and usable (web enabled, data warehouse, online reports)

7. Abbreviations

ABS	Australian Bureau of Statistics
ASCED	Australian Standard Classification of Education
ATSI	Aboriginal and Torres Strait Islander
CEQ	Course Experience Questionnaire (survey)
CGS	Commonwealth Grant Scheme
CIHI	Canadian Institute for Health Information
DEST	Department of Education, Science and Training
DESTPAC	DEST's Student Reporting Validation Software Package
FOE	Field of Education
GDS	Graduate Destination Survey
HECS	Higher Education Contribution Scheme
HR	Human Resources
IAF	Institution Assessment Framework
ID	Identification (code)
IT	Information Technology
KPI	Key Performance Indicator
SELT	Student Experience of Learning and Teaching (survey)
TAC	Tertiary Admission Centre

8. References

- Arts**, Danielle, de Keizer, Nicolette and Scheiffer, Gert-Jan. *Defining and Improving Data Quality in Medical Registries*, Jun 2002.
- Bushell**, Sue. *Cleaning Up Your Act*, July 2003.
- CIHI**. *Data Quality Framework*, Ottawa, Ontario, 2003.
- DEST**. *Learning and Teaching Performance Fund Issues Paper*, April 2004.
- English**, L.P. *Improving Data Warehouse and Business Information Quality*. John Wiley & Sons, New York, 1999.
- Haebich**, W and Bowles, S and Associates. *A Methodology for Data Quality Management*, Feb 1998.
- Mamonski**, Jon. *Managing student relationships: keeping the customer satisfied*, June 2004.
- Loshin**. *Issues and Opportunities in Data Quality Management Coordination*, in DM Review magazine, April 2004 issue.
- Paua** Interface Ltd. *Data Quality Frameworks in Health*, report for NZ Health Dept, Feb 2004.
- Pipino**, L, Lee, Y and Wang, R. *Data Quality Assessment*, in Communications of the ACM, April 2002/Vol 44 Number 4.
- Rothenberg**, J. *Metadata to Support Data Quality and Longevity*, first IEEE Metadata Conference, April 1996.
- Strong**, D, Lee, Y and Wang, R. *Data Quality in Context*, in Communications of the ACM, May 1997/Vol 40 Number 5.
- Watson**, Liz. *2004 Enrolment Review Report*, March 2004

9. About the Authors

Andrea Matulick

After graduating in Electrical and Electronic Engineering and Computer and Mathematical Science in 1980, Andrea worked with the South Australian State Transport Authority from 1981 to 1988 as an Electrical Engineer and Railway Signalling Engineer. She was involved in developing a computer Signalling Failure Analysis System, and implementing Computerised Control of the Adelaide Metropolitan Railway system. After a break to start a family, Andrea was employed by the University of South Australia in the Planning Office as Information Services Support Officer, Information Analyst and DEST Data Coordinator from 1997 to 2002. Since 2002 she has relocated to the University of Adelaide as Management Information Analyst in the Office of Planning and Quality.

Lachlan Murdoch

Lachlan graduated with a Bachelor of Science degree in 1982 and subsequently worked in the wine and pharmaceutical industries for 4 years in market analyst positions. After moving to South Australia in 1986 he worked as Market Analyst in private industry, and then took up the role of Clinical Trials Data Coordinator at the University of Adelaide. When the funding ran out in 1989 he moved to the South Australian Institute of Technology as Planning Officer just in time for the Institute to amalgamate with the SACAE and become the University of South Australia. Since that time he has remained in a planning and management information role at the University and is currently Coordinator: Analysis and Strategy within the Planning and Assurances Services Unit.