Notes on preparing the UCT Student Application Data 2006-2014

Alex Montgomery

May 2015

DataFirst

1 Available Files

The datafiles were made available to DataFirst as a group of Excel spreadsheet documents pulled from an SQL database by ICTS. The original SQL datafiles are summarised in Figure 1. Many of the peripheral datafiles in this figure are effectively value labels that serve to match codes to values. This made it possible to condense the initially large amount of datafiles in the dataset to something more manageable.

The linking of these files was performed over phases as summarised in several tables below. The emboldened text at the top of each column reflects the name assigned to the agglomerated datafiles generated. Some tables were excluded from the final datafiles because they were unsalvageably messy and low on information (work experience datafile) or irrelevant (supporting document type datafile).

The final dataset contains a person level datafile, an applications form level datafile, a school subject level datafile and a tertiary education "fact" level datafile (each row partially details some respondent's previous tertiary education experience). A person level datafile with private information was also included in the initial version of the dataset.

Table 1: Phase one linkages

1.1 Housing Information	**1.2 Supporting Document Information**	**1.3 Tertiary Education Information**	**1.4 Secondary Education Information**
housing_offer_fact housing_dim housing_offer_data_dim housing_offer_dim	document_fact document_type_dim	tertiary_education_fact tertiary_education_dim	secondary_education_dim organisation_group_dim organisation_location_dim education_dim ext_academic_level_dim

Table 2: Phase 2 Linkages

2.1 Secondary Education Information Expanded
1.4 Secondary Education Information secondary_education_fact school_subject_dim

Table 3: Phase 3 Linkages

Application Datafile	Person Datafile	Secondary Education Datafile	Tertiary Education Datafile
df_applicant_fact	df_bio_demo	1.4 Secondary Education Information	df_tertiary_education_fact
df_academic_plan_dim	df_key_subjects_fact	secondary_education_fact	df_tertiary_education_dim
df_academic_program_dim		school_subject_dim	
df_admissions_decision_dim			
df_applicant_data_dim			
df_applicant_info_dim			
df_nbt_performance_dim			
df_scholarship_dim			
df_term_year_dim			

2 Fixing Shifted Columns in Individual Level Datafile

Upon inspection of the individual/person level datafile it appeared that there was some column alignment skewing for some observations (a product of the original import by ICTS). To illustrate, see Table 4. Observations 5, 7 and 8 are shifted to the right.

It proved difficult to systematically identify the observations for which this was the case because many of the fields were missing for variables that were far to the right and had easily identifiable characteristics. As a result, qualitative differences between observations for the same variable were not always clear which made the shifts hard to find. To illustrate, there was an e-mail address variable in one of the rightmost columns. We would expect to see a "@" symbol as part of the string. If the "@" is not present in one of the entries for this variable it would be a good indication that the entry was qualitatively different and that the entries for that observation might have been shifted to the right. If all observations had non-missing entries for something like an e-mail address type variable it would be easy to figure out how far right the cells for each observation had shifted. Unfortunately, this was not the case.

It was possible, however, to look over enough of these shifted observations which (by looking at variables that were more "central" - the tradeoff here is that one wouldn't pick up all the shifted observations), when placed alongside one another, seemed to exhibit a discernible pattern. The shifting problem was seemingly caused by spaces (measured in cells) between two sets of open inverted commas for the same row observation. It turned out that for each cell between an opening set of inverted of inverted commas and a closing set observations would be shifted that many cells to the right. This was fixed manually in the original excel file as there were only 10 observations with this issue.

Figure 1: Diagram of related datafiles

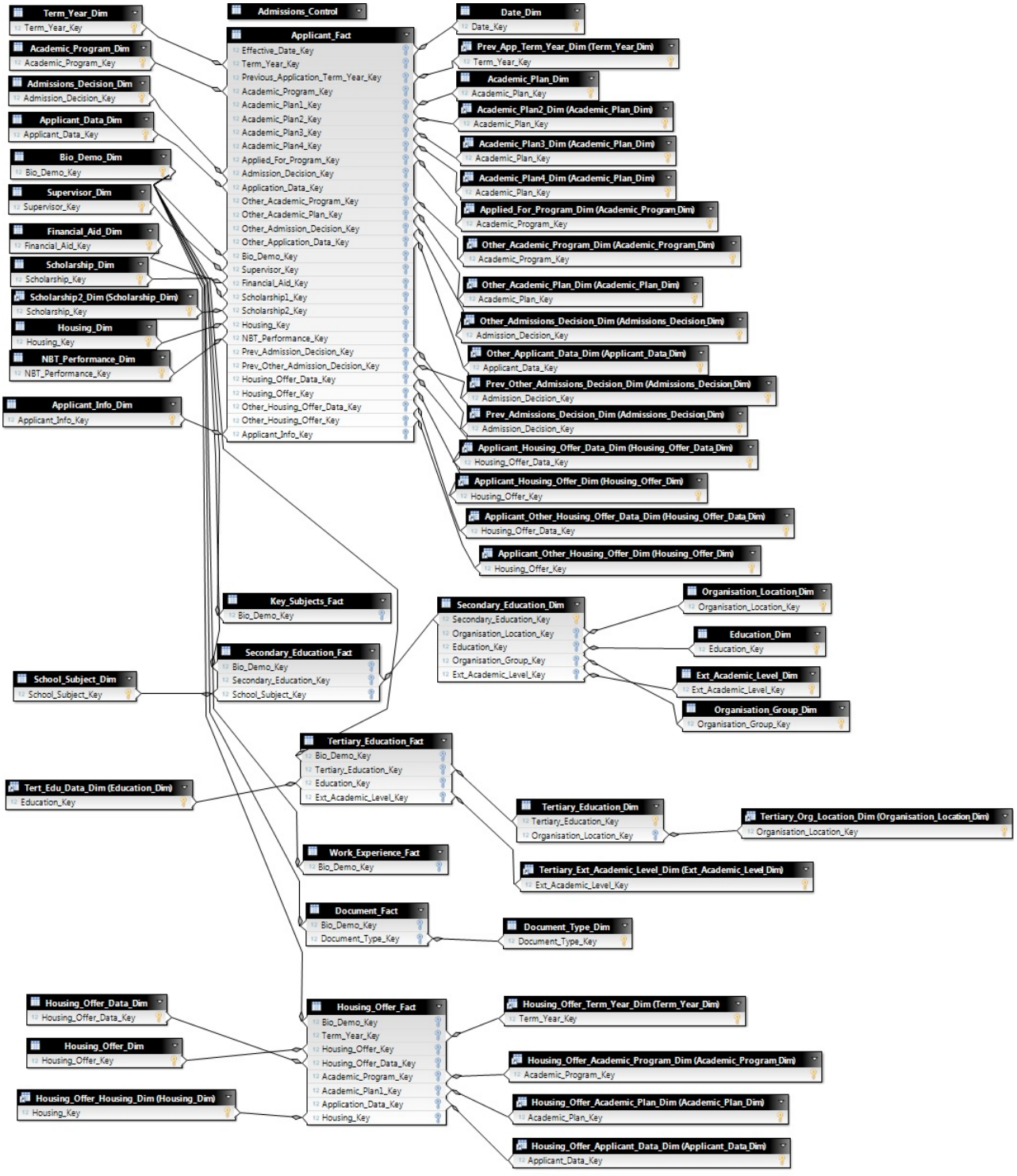

Table 4: Cell shifting

n	As and Bs	Random Number	Another random number	1s and 2s	Gs and Hs	Outside File?
1	A	0.19	0.04	1	G	
2	B	0.73	0.62	1	H	
3	A	0.97	0.97	2	H	
4	B	0.71	0.35	2	G	
5	A		0.06	0.38	1	G
6	B	0.26	0.83	1	G	
7	B		0.82	0.66	2	H
8	A		0.31	0.60	2	H
9	A	0.96	0.05	2	H	
10	A	0.21	0.52	2	H	