Classifying Australian PhD Bibliographic Thesis Records by ANZSRC Field of Research Codes

Report on a Study for the Research Excellence Branch, Australian Research Council

Final Report

July 2011

Peter Macauley (RMIT University), Terry Evans (Deakin University) & Margot Pearson (Australian National University)

Table of contents

1. Introduction

The *Classifying Australian PhD Bibliographic Thesis Records by ANZSRC Field of Research Codes* project, funded by the Research Excellence Branch of the Australian Research Council, involves the provision of three tasks:

1. a copy of the database of PhD thesis records for the period 2007 to 2009, coded by up to three ANZSRC (Australian and New Zealand Standard Research Classification) Fields of Research (6-digit level);

2. a copy of the database of any previously un-coded PhD records, where available, for the period prior to 2007, coded by up to three ANZSRC Fields of Research (6-digit level); and

3. provision of a brief report detailing the methods and approach used to classify the 2007-2009 theses by ANZSRC Fields of Research codes and any recommendations.

This report satisfies the third requirement. This project extends a project entitled *Classifying Australian PhD Theses by Research Fields, Courses and Disciplines* undertaken by the authors for the Research Excellence Branch, Australian Research Council (Macauley, Evans & Pearson, 2009). It also relates to two Australian Research Council Discovery Projects: *Research capacity-building: the development of Australian PhD programs in national and emerging global contexts* (Evans, Macauley and Pearson); and *Australian doctoral graduates' publication, professional and community outcomes* (Evans and Macauley). These research projects each partly involved coding the bibliographic records of Australian PhD theses. However, where the previous Research Excellence Branch and Discovery Grant projects differed from the current project is that the previous projects were coded by the Australian Standard Classification of Education (ASCED) (ABS 2001) and/or the Research Fields, Courses and Disciplines (RFCD) classifications, whereas this project adopts the new ANZSRC schema (2008). Furthermore, whereas the previous projects allocated a single code per thesis record, the current project allocates up to three codes per thesis record. This approach mirrors what which applies to research grant and publication coding. Consistent between these four projects is that the database of Australian PhDs was constructed from downloaded bibliographic records from the National Bibliographic Database, Libraries Australia. The projects mentioned above involved downloading bibliographic records of all PhD theses from Australian universities from Libraries Australia.

For this project, PhD records were downloaded from the Libraries Australia catalogue in a format which enabled importation into an Excel spreadsheet. A complex search strategy was constructed for the previous projects to determine the relevant records for downloading. This search strategy was used again for the current study. Once in the spreadsheet, the records where sorted, checked, and any duplicates or false drops were removed. Seven people were employed to code the records and, where possible, the records were distributed to coders according to their expertise. The seven coders chosen for the project demonstrated a wide range of relevant expertise between them. Up to three ANZSRC codes were allocated to each of the 9051 thesis records downloaded which will enable further bibliometric analyses of the 53,715 records provided in the 1987-2006 database supplied to the Research Excellence Branch in 2008. The result is the most comprehensive coded database of Australian PhD thesis records available.

The conclusions identified in this report have two broad themes relating to: the allocation of up to three ANZSRC codes (at six digit level) for each thesis record; and the current issues relating to thesis records no longer being submitted to a central catalogue or repository.

As we pointed out in our previous report, 'it is important to note that this database constitutes a valuable research resource in its own right. It provides an alternative source of data about research training with a focus on research output and research capacity building rather than input as does data on enrolment. The database is significant as it can be used to track knowledge production in Australia …' (Macauley, Evans & Pearson, 2009, p. 3).

2. Method and approach

Rationale

This project involved the provision of a database of PhD thesis records for the period 2007-2009, plus any previously un-coded records coded by ANZSRC Fields of Research, and a report detailing the methods and approach used to classify the theses by ANZSRC Fields of Research codes and any recommendations.

Bibliographic records from Libraries Australia

The new database of Australian PhDs (2007-2009), like its predecessor (1987-2006), has been constructed from downloaded bibliographic records from the National Bibliographic Database, Libraries Australia. Additionally, to ensure the most comprehensive coverage, where possible, individual library catalogues from Australian universities have been searched and any records not listed on Libraries Australia have been included. Importantly, during the current project it became clear that many university libraries only upload their PhD theses records and/or e-theses to their institutional repositories and, it appears, many thesis records are no longer provided to Libraries Australia. This means the database provided as part of the project is less comprehensive than the previous database and than was expected when the project was commissioned.

In addition to the initial searches for the original database, the National Library provided quarterly updates of new bibliographic records of Australian PhD theses uploaded from the respective university libraries into its national database. This report focuses upon the period 2007-2009 and includes 9051 coded PhD records. The earlier report covered the years 1987-2006 and was based on the analysis of 53,715 records for the two decade period.

Understandably there can be some variation of ‘publication’ years of theses occur which can marginally affect the number of PhD theses counted for a particular year. This can cause a ‘slippage’ from one year to another due to differing interpretations of the publication year as PhD theses are manuscripts and, as such, are not technically published. In many cases, libraries consider the publication date to be the thesis submission date, while others use the date of doctoral confirmation from the academic board or senate and some may use the date of graduation. This slippage can commonly result in the publication date differing from official university reporting statistics by one year.

The Libraries Australia search strategy

To enable the relevant bibliographic records to be downloaded from Libraries Australia, an extremely complex search strategy was constructed. In the previous PhD coding projects the search strategy was modified a number of times to ensure we were finding the greatest number of relevant PhD records and reducing the number of false drops. This has been a very challenging task as differing interpretations of the Anglo American Cataloguing Rules by individual libraries and librarians can result in valid records not being picked up in the searches. Hence the reason for the strategy being revised a number of times. A result of these cataloguing inconsistencies is that we cannot categorically state we have located every PhD thesis record produced from Australian universities. If libraries were not cataloguing theses and/or not uploading the bibliographic records to the respective online catalogues, the records will remain invisible. The last issue has been exacerbated by some libraries deciding not to upload their thesis records to Libraries Australia and relying only on submission of records to their own institutional repository. In doing so, this can severely restrict dissemination and their PhD graduates doctoral research.

ANZSRC codes and coding

The Australian and New Zealand Standard Research Classification (ANZSRC) was used to code the database of Australian PhD thesis records. The ANZSRC schema produced by the Australian Bureau of Statistics and Statistics New Zealand and released in 2008 enables both Research and Development activity and other activity within the higher education sector to be categorised. The ANZSRC replaced the RFCD classification (ABS, 1998). The newer ANZSRC classification scheme provides a more finely detailed description of research areas. That is, 1,238 Fields (six digit level) as opposed to 898 Subjects in the RFCD classification. It has 157 Groups (four digit level) compared 139 RFCD Disciplines and 22 Divisions (two digit level) rather than 24 RFCD Divisions in the older scheme. Both coding schemas are arranged in hierarchical structures. The categories in the classification include recognised academic disciplines and related major sub-fields taught at universities or tertiary institutions, major fields of research investigated by national research institutions and organisations, and emerging areas of study.

The coding procedures

The PhD thesis records were downloaded from the National Bibliographic Database, Libraries Australia, in bar delimited format which enabled us to import them into an Excel spreadsheet. Once in the spreadsheet, the records where sorted, checked, and duplicates and false drops were removed. While the search strategy was amended to reduce the irrelevant records manual checks of the downloaded records were still required.

Seven people were employed to code the records and, where possible, the records were distributed to coders according to their expertise. It should be noted that the coders used the bibliographic records produced by librarians from all Australian universities rather than coding directly from the actual theses. The ANZSRC code allocated to each thesis record was judged on a number of factors including: the thesis title, subject headings and call numbers (allocated by the institution’s librarians), the Department/School/Faculty, and an abstract (where provided). Additional resources were used to clarify terms including specialist print and online dictionaries, and connecting online to Libraries Australia for relevant links. Wikipedia was an excellent source of information, particularly when searching for definitions, with coders gaining confidence in the utility of its contents. To ensure consistency a number of processes were implemented. All coders were provided with training and a pairing system was initiated for newer coders to be partnered with a more experienced coder. While there were some face-to-face meetings, most of the communication took place via group email with all coders being involved. Any urgent issues were resolved over the phone between CI Macauley and the coder concerned.

The coders and their expertise

The seven coders chosen for the project demonstrated a wide range of relevant expertise between them. This is shown in Table 1 below. While one could not expect seven people to be expert in all areas, their expertise covered many disciplines. If a coder felt they were unable to code records in particular fields, they were referred to another coder. Some of the coders are highly successful ‘trivia competition’ players which requires them to have a good breadth of knowledge. While specialist knowledge is certainly the best option, realistically, there is no financial or practical way to employ a group of specialists to cover 22 Divisions, 157 Groups, and 1238 Fields of Research! It is no surprise that the average age of the coders was 53 years, as life experiences enhance the skills and knowledge of those individuals. Importantly, five of the seven coders had formal qualifications in cataloguing and classification—that is, they are librarians—which ensure they are trained to seek information and accurately represent that information through national and international schemas.

Table 1: Qualifications and Expertise of Coders

Coder	Qualifications	Areas of Expertise	Career History	Age	Gender
1	BSc. (Hons) Metallurgy, MApp Sci. Metallurgy, MA Librarianship	Physical Sciences, Engineering and Technology, History (especially History of Science and Technology and Maritime History), Plant Sciences and Biology, Horticulture	Metallurgist, Lecturer in Metallurgy/Materials Science and History of Engineering, Librarian	70	Male
2	BMus (Hons), BA, M. Info Mgt	Music, Information Management, History, Media, Journalism	Librarian, Media Researcher, Research Assistant, Archivist, Classical Musician	33	Female
3	BA, Grad. Dip. Lib.	Science, Philosophy, Languages, Librarianship	Laboratory Assistant, Cataloguer	69	Female
4	BSc, Grad. Dip. Dietetics, Grad. Dip. Arts (Lib & Info Studies)	Dietetics, Health, Science, Information Management	Dietician, Research Fellow, Librarian	53	Female
5	BA	History, Australian Studies	Real estate, Research Assistant	54	Male
6	BA; GradDipAppSc (Lib & InfoMgt); MAppSc (Lib & InfoMgt)	History, Psychology, Criminology, Information Management	Librarian, Editor	26	Female
7	AALIA	Generalist	Librarian for 46 years. Specialist thesis cataloguer for 30 years	64	Female

3. Findings related to the method and approach

Number of records coded

For this report, a total of 9051 PhD records were coded (see Figure 1 below) for the years 2007 to 2009 inclusive.

Figure 1: Number of PhD Thesis Records by Year

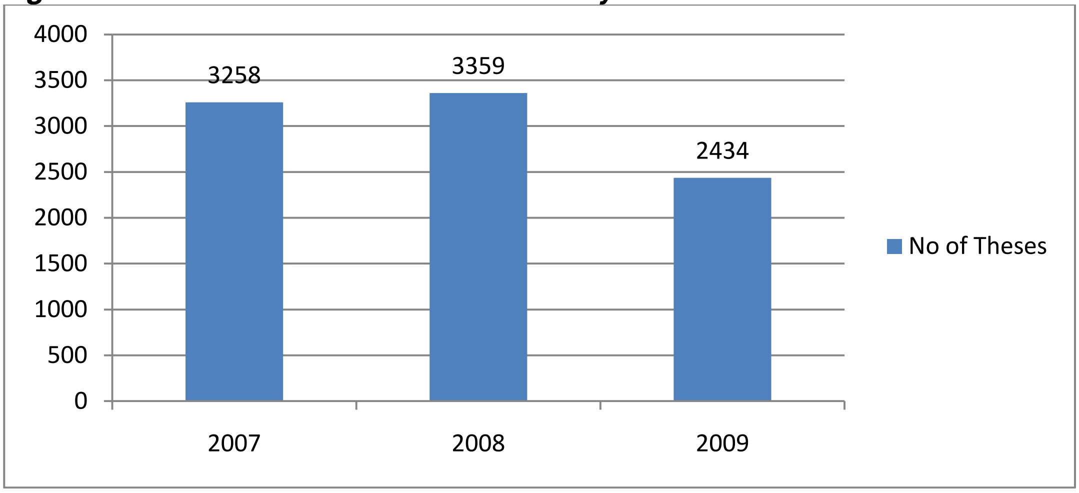


Figure 1 shows that a total of 9051 records were coded: 36.0% were from 2007; 37.1% from 2008, and 26.9% from 2009. It is not surprising that fewer records come from 2009 as there are often delays in cataloguing and uploading the bibliographic records to Libraries Australia.

Comparison of PhD thesis records in the database with number of ‘doctorate by research’ graduates reported to the Department of Education, Employment and Workplace Relations (DEEWR)

A significant limitation of this project was the consistent decline in the number of PhD thesis records being uploaded to Libraries Australia in comparison with the number of ‘doctorate by research’ completion figures published by Department of Education, Employment and Workplace Relations (DEEWR). Table 2 shows the number of doctorate by research graduates according to the figures from DEEWR (2011) compared with the number of PhD theses from the database for the corresponding years. It should be noted that the DEEWR figures are for doctorate by research graduates, consequently all research based professional doctorates are included in that figure. If the professional doctorate completions were excluded from the DEEWR statistics, the percentage of available PhD thesis records to graduates reported would increase, probably by 2-4%. While we, or DEEWR, cannot be definitive about the numbers of PhD graduates, it is likely that the coverage of PhD thesis records in our database is more comprehensive than Table 2 suggests due to this report and associated database being limited to PhDs only. It is also important to note that due to the slippage, mentioned previously, and inconsistencies of reporting the ‘publication’ years of theses, the data for each year is not fully comparable. Table 2 below shows the number of coded records we have reported in the previous report (Macauley, Evans & Pearson, 2009) and the records for the years 2007-2009 discussed in this report. The comparison cannot be made for earlier years as those figures included all higher degree by research graduations (i.e. including masters by research).

Table 2: Comparison of PhD thesis records in the database with number of doctorate by research graduates (DEEWR, 2011)

Year	PhD thesis record count	Doctorate by research graduates	Percentage of PhD thesis records to graduates
1991	1478	1519	97.3%
1992	1687	1522	110.8%
1993	1842	1793	102.7%
1994	2065	2201	93.8%
1995	2501	2437	102.6%
1996	2798	2905	93.1%
1997	3262	3346	97.5%
1998	3225	3446	93.6%
1999	3469	3665	94.7%
2000	3552	3793	93.6%
2001	3624	3933	92.1%
2002	3873	4295	90.2%
2003	3971	4722	84.1%
2004	4071	4900	83.1%
2005	3672	5244	70.0%
2006	3286	5519	59.5%
2007	3258	5721	56.9%
2008	3359	5786	58.1%
2009	2434	5796	42.0%

Table 2 shows, since 2005, the uploading of PhD bibliographic records to Libraries Australia has steadily declined affecting the comprehensiveness of the National Bibliographic Database. A possible cause of this has been a shift in university priorities.

Issues in cataloguing, uploading and coding

In the previous study (Macauley, Evans & Pearson, 2009) we found a small number of university libraries were tardy in uploading their PhD records to Libraries Australia. In this study, two years later, it is obvious that this problem has increased. Some libraries across the Group of Eight, the Australian Technology Network, the Innovative Research Universities, and the unaligned universities have become extremely patchy and severely overdue with their PhD record uploads. This is an issue of great concern for those who value and use a national record of PhD theses for their work.

While cataloguing delays are perhaps inevitable, this possibly exacerbated by theses requiring original cataloguing of the document. This requires more time and skill than the 'copy-cataloguing' of books, for example, where bibliographic details are provided on the imprint page.

Limits to coverage in all years include:

- Some theses may never have been lodged in the appropriate library
- Some theses may never have been catalogued (i.e. lost in the system)
- Mistakes in cataloguing, for instance cataloguing a PhD thesis as a masters thesis, will mean the bibliographic records will not be picked up by our search strategy.

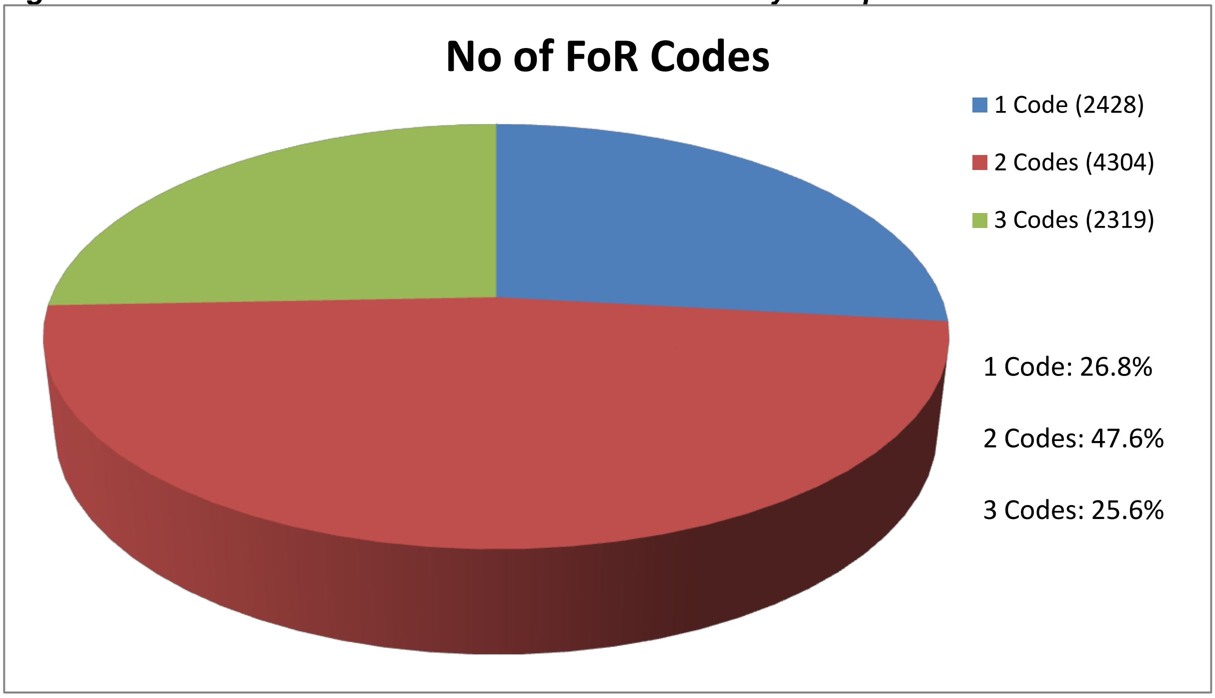
The project made a number of checks to compare institutional catalogues with the Libraries Australia records and to download missing records found for coding. In some cases this proved problematic. Using institutional public access catalogues did not always enable searches to be undertaken by 'thesis'. Given the value of such a comprehensive database for researchers and doctoral candidates it is important that regular, consistent and accurate uploading of PhD records to Libraries Australia is sustained or some other option put in place.

Allocating up to three codes.

Unlike the previous coding projects where one code was allocated to each of the PhD bibliographic records, this project allocated up to three codes to each of the records. The allocation of up to three ANZSRC codes is a requirement of many other research activities in universities. For example, in submitting Australian Research Council and National Health and Medical Research Council Grant applications, and for publications submitted to institutional repositories for DEEWR reporting purposes. To provide a more consistent indication of the relevant fields of research for each thesis and a more accurate indication of Australian 'research training' output and future research capacity, it was decided to allocate up to three codes for each thesis. Arguably, the most suitable people to allocate codes are the candidates or graduands themselves (and their supervisors), however, this has not been required within universities and so such coding rarely forms part of a thesis record for the period up to 2009. Comments from coders for the previous projects suggested that, at times, restricting a thesis to one code was difficult and allocating multiple codes would provide a more appropriate description of the research conducted and findings produced.

Subsequent feedback from this project reinforced those earlier views. Allowing up to three codes provided more flexibility, as well as a better indication of the content of a thesis and, arguably, the graduate's research capabilities. While previously some coders agonised over which code to allocate, they could now simply use the two or three they were considering. In many cases, this made for easier coding; however, decisions still had to be made regarding which code was named first, second or third. For the purposes of this study and comparative analysis with the original single-coded database, the coders were advised that their first named code would count as the primary Field of Research. Unlike, research grant and publication coding, no attempt was made to add percentages to each code as only the PhD authors or their supervisors would be equipped to do so. The coders who had worked on the previous studies suggested that, while it depends on the knowledge of the coder, using multiple codes provided a better and more accurate outcome. They also spoke, however, of the occasional difficulty of differentiating between primary and secondary codes. Generally, the more they knew about a thesis topic, the more likely they were to allocate it more than one code. Figure 2 below shows the breakdown of the number of codes used for thesis records.

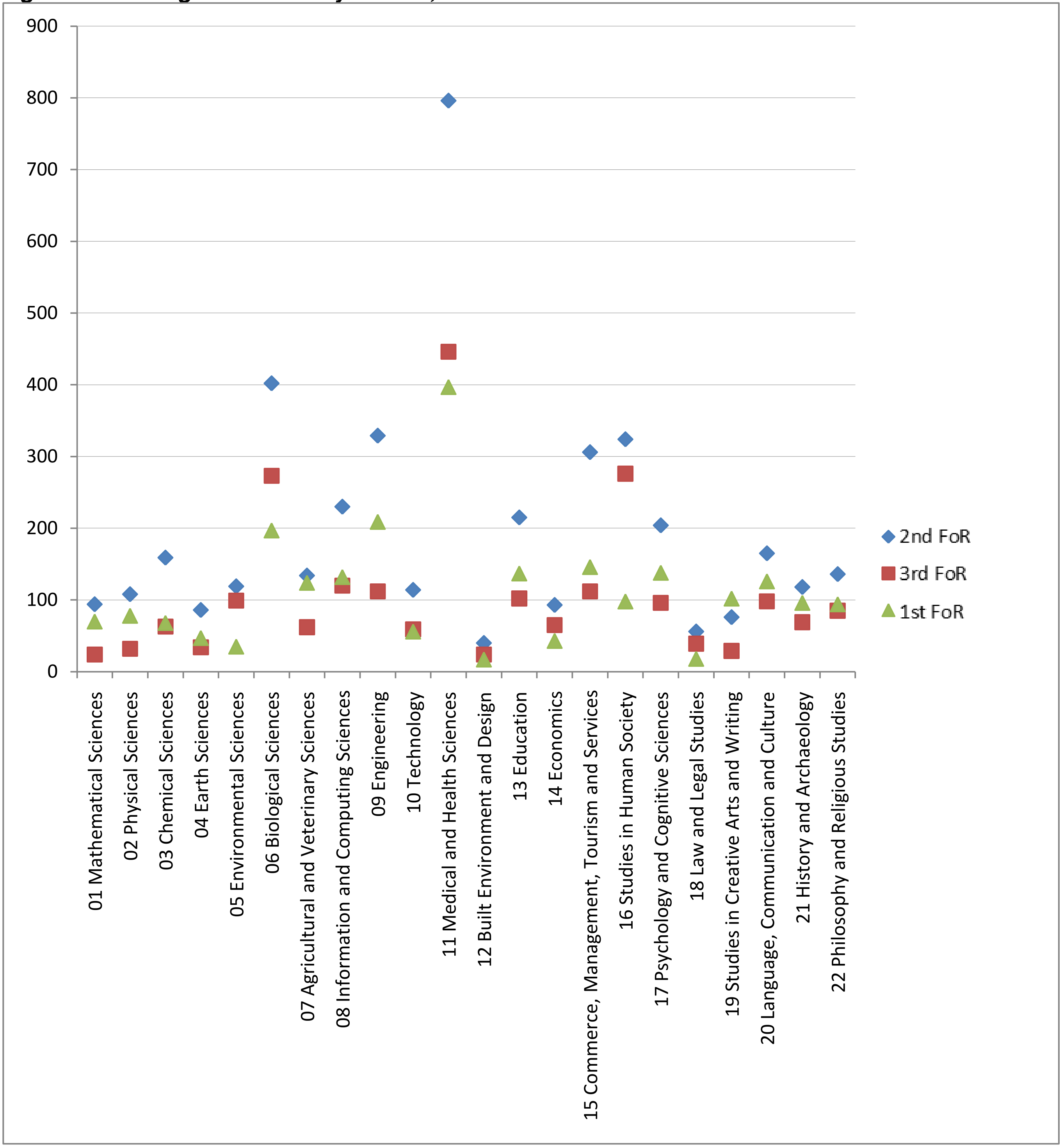
Figure 2: Numbers of PhD Theses Records Classified by Multiple ANZSRC Codes



As can be seen in Figure 2, just over one quarter (26.8%) of the PhD records were allocated with (only) one ANZSRC code; roughly half (47.6%) had two codes; and around one quarter (25.6%) were given three codes. Advice from the coding team suggested that in many cases theses related to one code only so the choice was relatively easy; there was no need to look further afield. Allocating two or three codes was both a blessing and a dilemma. On some occasions it made the task easier and at other times, it made it more difficult, as mentioned above.

It is important to note that the allocation of some thesis records to all of the three coding options occurred across the Divisions, as shown in Figure 3 and Table 3. All Divisions had thesis records coded with one, two or three FoR codes, from Medical Sciences with the largest number of thesis records coded (1639) through to the Division with the smallest number of thesis records coded - Built Environment and Design (81).

Figure 3: Two Digit Divisions by 1st FoR, 2nd FoR & 3rd FoR



The pattern across the three coding options varied within Divisions as can be seen more clearly in Table 3. At the extremes 12 Divisions had more thesis records than the average (26.83%), with one code only and eight Divisions had more thesis records than the average (25.62%) with three codes allocated. Those with the largest percentage with three codes allocated were Environmental Sciences and Studies in Human Society, both divisions where less than a quarter of the records were given one code only (13.83% and 14.04% respectively), and where more inter/multidisciplinary research is expected to be being carried out. However all Divisions bar Studies in Creative Arts and Writing (49.28% one code only), had the clear majority of their thesis records given two or more codes, indicating the potential for using a more flexible coding system. As the coders found the provision of up to three codes gives a much richer description of the content of theses. Thus the use of multiple codes can provide a much more detailed overview of knowledge production in Australia through PhDs.

Table 3: Coding Within Divisions

Division	FoR1	Within Div FoR1 %	FoR2	Within Div FoR2 %	FoR3	Within Div FoR3 %	Total
01 Mathematical Sciences	70	37.23%	94	50.00%	24	12.77%	188
02 Physical Sciences	78	35.78%	108	49.54%	32	14.68%	218
03 Chemical Sciences	68	23.45%	159	54.83%	63	21.72%	290
04 Earth Sciences	47	28.14%	86	51.50%	34	20.36%	167
05 Environmental Sciences	35	13.83%	119	47.04%	99	39.13%	253
06 Biological Sciences	197	22.59%	402	46.10%	273	31.31%	872
07 Agricultural and Veterinary Sciences	124	38.75%	134	41.88%	62	19.38%	320
08 Information and Computing Sciences	132	27.39%	230	47.72%	120	24.90%	482
09 Engineering	209	32.15%	329	50.62%	112	17.23%	650
10 Technology	56	24.45%	114	49.78%	59	25.76%	229
11 Medical and Health Sciences	397	24.22%	796	48.57%	446	27.21%	1639
12 Built Environment and Design	17	20.99%	40	49.38%	24	29.63%	81
13 Education	137	30.18%	215	47.36%	102	22.47%	454
14 Economics	43	21.39%	93	46.27%	65	32.34%	201
15 Commerce, Management, Tourism and Services	146	25.89%	306	54.26%	112	19.86%	564
16 Studies in Human Society	98	14.04%	324	46.42%	276	39.54%	698
17 Psychology and Cognitive Sciences	138	31.51%	204	46.58%	96	21.92%	438
18 Law and Legal Studies	18	15.93%	56	49.56%	39	34.51%	113
19 Studies in Creative Arts and Writing	102	49.28%	76	36.71%	29	14.01%	207
20 Language, Communication and Culture	126	32.39%	165	42.42%	98	25.19%	389
21 History and Archaeology	96	33.92%	118	41.70%	69	24.38%	283
22 Philosophy and Religious Studies	94	29.84%	136	43.17%	85	26.98%	315
Totals	2428	26.83%	4304	47.55%	2319	25.62%	9051

Thesis records coded as ‘Not Elsewhere Classified’

A complicating factor in identifying patterns of growth in Divisions, Groups and Fields occurs where there are a high number of the ‘not elsewhere classified’ (NEC) codes used. High numbers of NEC codes can indicate that revised or new codes may be needed for a field. This is not surprising where PhDs are concerned as they are required to represent original and substantial contributions to knowledge and, consequently, may expand the boundaries (and their classifications) of knowledge. However, if NEC codes are used frequently this can lead to ambiguous reporting of research and an under-representation in particular subjects, disciplines and divisions. This was a major issue raised in the earlier report, *Classifying Australian PhD Theses by Research Fields, Courses and Disciplines* (Macauley, Evans & Pearson, 2008, pp. 16-17). Feedback from the coders for this project, suggested this was less of an issue for the 2007-2009 PhD records which have been coded using the ANZSRC schema which was introduced in 2008 compared, with the earlier RFCD classification which was released in 1998. Their observations were backed up by the data. Only 292 records (using the first code only for comparative purposes) were coded to NEC. This amounts to 3.2% of the 9051 records coded in this project. This figure, using the more comprehensive 2008 ANZSRC coding schema, compares extremely favourably to 4360 NEC coded records (8.1%) of the 53,715 records coded in the previous project using the 1998 RFCD schema. The reduction of more than 50% of the NEC coded records is probably due to the more comprehensive nature of the ANZSRC schema with an expansion of six digit fields from 898 to 1238 (38%) which included the addition of new Fields of Research not previously identified in 1998 and a finer degree of specificity in some areas. However, there were still some disciplines which were problematic.

In the 2009 report, covering the period 1987-2006, the six most frequent number of NEC thesis records classified by RFCD subjects are listed in Table 4 below.

Table 4: Most Frequent number of thesis records not elsewhere classified (NEC) in RFCD subjects, 1987–2006

RFCD Subject	No	% of Total NEC Coded Records
Education Studies not elsewhere classified	584	13.4
Nursing not elsewhere classified	168	3.9
Biochemistry and Cell Biology not elsewhere classified	153	3.5
Genetics not elsewhere classified	137	3.1
Literature Studies not elsewhere classified	131	3.0
Business and Management not elsewhere classified	109	2.5

Table 5 (below) lists the most likely NEC codes for the 2007-2009 thesis records coded by ANZSRC.

Table 5: Most Frequent number of thesis records not elsewhere classified (NEC) in ANZSRC fields, 2007-2009 1st Listed FoR)

ANZSRC Field	No	% of Total NEC Coded Records
Specialist Studies in Education not elsewhere classified	28	9.6
Nursing not elsewhere classified	21	7.2
Business and Management not elsewhere classified	13	4.5
Biochemistry and Cell Biology not elsewhere classified	9	3.0
Genetics not elsewhere classified	9	3.0
Sociology not elsewhere classified	9	3.0

The two lists are remarkably similar and even though the overall numbers of NEC records has declined using the new schema, the same Fields of Research appear to be most challenging to code. This suggests that some more refining of classifications may be required.

As mentioned above, there were 292 of the first listed (primary) FoR coded thesis records to NEC. Additionally, there were 346 second listed, and 133 third listed FoR coded thesis records to NEC. Once again, there was consistency with those coded to particular NEC Fields of Study. For the second listed FoR, 0604 Genetics, was the most frequent code, followed closely by 1503 Business and Management and 1303 Specialist Studies in Education. With regard to the third listed FoR, 0604 Genetics, was the most frequent code again followed by 1303 Specialist Studies in Education.

4. Submission of PhD Bibliographic Records to Libraries Australia and Repositories

The greatest challenge for this project was accessing the thesis records of recently completed PhD theses. The format of PhD theses records and the nature of storage has changed significantly in recent years. In particular, the shift from open access library catalogues to password protected repository storage means that many thesis records (and e-theses) are no longer made publicly available. There is no longer a central repository or catalogue for Australian PhD records which is actively supported by all Australian university libraries. While Libraries Australia is a subset of the Trove interface implemented in November 2009 by the National Library (see http://trove.nla.gov.au/) it does not include all completed Australian PhD records. Some Australian universities do not upload their thesis records from their repositories, and uploads of thesis records to Libraries Australia has decreased over recent years. This has reduced the number of PhD records available for downloading for coding in this project. More importantly, this adversely affects the comprehensiveness of the Libraries Australia catalogue, and consequently, public access to the information.

Currently, there are seven different types of software used for university open access repositories in Australia. The functionality and accessibility for locating PhD thesis records range from closed to open. Some interfaces are extremely poor with no direct means of searching for theses. It is obvious some of these interfaces are designed more for accessing journal articles rather than other media. Nevertheless, there are some excellent, well designed, interfaces where searches can not only be refined to search for theses, but also can searched by ANZSRC codes and by type of thesis (i.e. PhD, professional doctorate, masters by research). One university's interface even enables searches by the format of thesis such as 'traditional', 'by publication' and 'by creative works'. This repository also enabled downloading of records in a variety of formats and, from the perspectives of this project and of public access more broadly, this repository is exemplary. As most of the institutional repositories are harvested by major search engines, including Google Scholar, having genuine open access is a very sensible strategy if the goal is to make a university's publications, theses etc known to the world.

It is perhaps ironic that, while access to the research output of Australian university researchers, including PhD scholars, is potentially greatly enhanced through the open access repositories, in many instances access has been effectively closed to these scholars. The decrease in thesis records being uploaded to the National Bibliographic Database—Libraries Australia—has also reduced the comprehensiveness of this database. Due to the nature and structure of the respective repositories, downloading the majority of the bibliographic records of PhD theses is either impossible or very difficult and time consuming, which seriously affected the number of records accessed for this project. Contrary to the original intentions, the decommissioning of the Australasian Digital Theses Program and its incorporation into the National Library of Australia has exacerbated the problem. While all Australian universities have institutional repositories that, in theory, are harvested by the National Library of Australia's Trove Discovery Service, if thesis bibliographic records are not loaded into the repositories and if the interrogation of the repositories is not possible, Trove is unable to find and harvest those resources. Therefore, people using Trove to find Australian theses will not be able find such theses.

5. Summary and conclusions

This project is derived from a previous project entitled *Classifying Australian PhD Theses by Research Fields, Courses and Disciplines* undertaken by the authors for the Research Excellence Branch, Australian Research Council (Macauley, Evans & Pearson, 2009). It is also derived partly from two Australian Research Council Discovery Projects conducted by the authors. These projects each used coding the bibliographic records of Australian PhD theses for part of their data collection. A significant element of the current project is that it used multiple (up to three) ANZSRC coding whereas the earlier projects used single ASCED or RFCD coding.

The project demonstrates the utility of multi-coding PhD thesis records in order to reflect the breadth of research contributions of theses and, by implication, the breadth of research capacities embodied in the graduates. Using independent coders, 74% of theses were shown to contribute to two or three fields. Doctoral candidates (with their supervisors), arguably, are likely to see even greater complexity in their theses and so one might expect that almost all theses would be multi-coded by them. From the perspective of both understanding doctoral research productivity and where it occurs, and of research workforce planning for the future, the use of multiple codes on thesis records in university libraries and repositories is a significant improvement. The project identified difficulties arising from the move to institutional repositories intended to improve public access to university research and scholarship as well as provide support for the implementation of an Australian research assessment—the Research Quality Framework which became Excellence in Research for Australia. It appears that the implementation of research repositories in all Australian universities has been done in a way that has, in many cases, either ignored PhD theses and their bibliographic records, or has unintentionally hidden them from public access. Whereas, previously, university libraries held all PhD theses and their catalogues were viewable publicly, and usually the theses were available to the public, although usually with some limitations.

This report leads to four main conclusions arising from both the bibliographic data and coding, and from the experiences of the conduct of the project itself, together with information obtained from persons involved with institutional repositories and higher degree by research administration and management. These conclusions are presented briefly below.

Conclusion 1

It is clear that, as Australian universities move to electronic lodgement of PhD theses in their institutional repositories that there are now differences in practices emerging. In some institutions (some) theses are no longer being lodged in their libraries and being catalogued by experienced cataloguing librarians. Rather, theses are being lodged electronically in their repositories. Whilst this is commendable in itself, an unintended consequence appears to be that the repository records are not of the usual cataloguing standard and thus, contain inadequate data for discipline coding purposes (for this project at least) and also reduces the searchable fields and consequently, accessibility to those theses. Furthermore, some institutional repositories are ‘dark’, that is closed to public access. Therefore, thesis records are unable to be harvested by Trove and are invisible to the researchers, PhD students and others. This is in stark contrast to the NLA library records that are publicly available. We suggest, therefore, that it is in the interest of the university community in particular, and the general community, that all Australian universities ensure their institutional repositories are harvested by Trove to enable open access to Australian PhD and other records.

Conclusion 2

Further to Conclusion 1, it is noticeable there is no longer a clearly identified central repository for Australian produced PhD thesis records and/or e-theses. The demise of the Australasian Digital Theses Program has exacerbated the situation as has the focus on universities lodging thesis records and/or full text e-theses only in their own repositories. Pragmatically, Trove is the logical avenue for centralising Australia’s thesis records. However, to ensure the quality and usefulness of the PhD records they need to be of international bibliographic cataloguing standard rather than the pared down records used for many repositories. This issue would be resolved if university libraries reverted to the standard practice of cataloguing their own institution’s PhD theses and uploading the records to Libraries Australia which, in turn, is harvested by Trove. Another important advantage of having full bibliographic records uploaded to Libraries Australia is that those records contain ‘holding’ statements which list the libraries that have the theses and the format of the versions held.

Conclusion 3

This Report shows that the use of up to three ANZSRC codes to denote the fields addressed by a PhD thesis is potentially very useful to understand the trends in both the production of doctoral research and also a proxy for future research capacity and its disciplinary locations. However, as noted above, universities otherwise laudable move to the lodgement of e-theses in repositories is producing a reduction in the number of professionally catalogued thesis records available through the NLA/Trove. It is concluded, therefore, that the matter of good quality coding of thesis records needs to be addressed both nationally and, subsequently, at institutional level. After discussion with persons involved in doctoral education, libraries and repositories the conclusion was reached that the solution would be for each university to require their PhD graduands, in consultation with their supervisors, to submit their theses to the university's repository and to allocate up to three ANZSRC codes which then form part of the permanent bibliographic record of the thesis. This has the advantage of the persons closest to the thesis (graduand and supervisor) undertaking the coding in what should then be the most accurate allocation of codes possible. This has the benefit of teaching graduands to allocate ANZSRC codes which is something that researchers normally have to do throughout their research careers, in particular, for their publications and research grant submissions. Furthermore, it reduces the work required by cataloguers in university libraries and repositories and assists in producing high quality cataloguing records.

Conclusion 4

Tables 4 and 5 above show that there remain some ANZSRC groups where there are disproportionately high numbers of theses allocated to their 'not elsewhere classified' fields. These groups are largely the same as the RFCD subjects where the previous report (Macauley, Evans & Pearson, 2009) also identified disproportionately high numbers of theses allocated to the 'not elsewhere classified' subjects. It may be concluded, therefore, that the Australian and New Zealand Standard Research Classification requires modification to include one or more new fields in the following groups.

- 1303 Specialist Studies in Education
- 1110 Nursing
- 1503 Business and Management
- 0601 Biochemistry and Cell Biology
- 0604 Genetics
- 1608 Sociology

Addressing the issues raised by these conclusions will require a number of stakeholders to contribute to achieving an optimal outcome. It will involve consultation with various bodies, e.g. CAUL, DDoGs, the NLA, and others, to enable the National Library to have robust and reliable records of all Australian PhD theses; involve candidates and supervisors in coding theses (and, thereby, provide the former with the skills to do such in the future); and lead to a further refinement of some ANZSRC codes in due course that would improve coding of the related areas.

6. Acknowledgements

The Chief Investigators would like to acknowledge the contributions of the following people:

> The seven research assistants who coded the records: Andrew, Clara, Donna, Lesa, Niki, Tony, and Peta who coordinated the majority of the coding; Dr Kevin Ryland who provided advice on coding and mapping codes; and Dr Siddhi Pittayachawan provided some useful statistical expertise.

The National Library of Australia is also gratefully acknowledged for the providing quarterly updates of the PhD bibliographic records and advice.

7. References

Australian Bureau of Statistics. (1998). *1297.0 - Australian Standard Research Classification (ASRC)*. Canberra.

Australian Bureau of Statistics. (2001). *1272.0 - Australian Standard Classification of Education (ASCED)*. Canberra.

Australian Bureau of Statistics, & Statistics New Zealand. (2008). *1297.0 - Australian and New Zealand Standard Research Classification (ANZSRC)*. Canberra.

Department of Education, Employment and Workplace Relations. (2011). *Award course completion 2009: selected higher education statistics tables*. Canberra. http://www.deewr.gov.au/HigherEducation/Publications/HEStatistics/Publications/Pages/Students.aspx

Macauley, Peter, Evans, Terry and Pearson, Margot (2009) *Classifying Australian PhD Theses by Research Fields, Courses and Disciplines: report on a study for the Research Excellence Branch, Australian Research Council*. [Unpublished report] RMIT University, Deakin University and the Australian National University. http://prodmams.rmit.edu.au/qjcu4phay2ia.pdf