Can we improve thesis data quality?

Gerald Q. Maguire Jr.

While trying to assess the use of English and Swedish titles for theses at the 1st and 2nd cycle and to understand the maximum length of these titles (as input to the design of the new cover), it became apparent that there are problems with the data being inconsistent, inaccurate, or missing. This document examines the possibility of improved quality control for data regarding theses at KTH.

There are two major places where data about 1st and 2nd cycle theses is recorded at KTH: DiVA and LADOK.

- LADOK has information about the title and alternative title of each thesis for which a grade has been assigned.
- DiVA should have the meta data for each thesis and at least an archive copy of the actual thesis.

However, a recent examination of data in these two sources reveals that there are some problems with the data quality. Not only are there inconsistencies between the titles in LADOK and DiVA, but this data may differ from the titles in the actual thesis (with error rates in the range of 2 to ~4%). Additionally, some (~1%) degree projects have grades reported in LADOK, but there is no meta data for the corresponding thesis in DiVA.

Additionally, there are some clear examples of people not following administrative policies (ranging from more students than permitted for a given type of thesis to incorrect covers, incorrect placement of the covers and lack of cover, and lack of English or Swedish abstracts (Note that since 2010, there should be an abstract or summary in each language).

Section 1 looks at the data that currently exists. Section 2 raises the question of whether one could check the quality of the LADOK data using the data from DiVA and implicitly check the DiVA data for completeness based upon the LADOK data. However, it concludes that there is not a simple way to connected the LADOK and DiVA records, as currently the KTH ID of authors is rarely present in the DiVA records. Section 3 tell how you can get the data and reproduce these results or do your own computations on the data.

While the data is useful for a number of purposes, such as estimating the maximum lengths of titles and subtitles for the design of the cover and detecting some systematic problems in processes (such as administrators pasting the same value into both fields in LADOK despite their actually being two different titles). **My overall conclusion is that as manual data entry is prone to error[*]; hence, (1) there is a need to automatically enter the data into both DiVA and LADOK and (2) there is a need for more systematic quality control of the data. Otherwise, it would seem that we have to expect an error rate of several percent.**

[*] Data and equation entry into spreadsheets was already pointed out as a problem by Ray Panko in 1998, see also his 2016 paper, "What We Don't Know About Spreadsheet Errors Today: The Facts, Why We Don't Believe Them, and What We Need to Do" https://arxiv.org/abs/1602.02601. His work points to a 1-5% human error rate, see: http://panko.com/HumanErr/SimpleNontrivial.html.

1 Status

To understand the thesis titles in preparation for the new cover that the GVS Communications unit is working on and to understand the usage of English and Swedish titles for the language committee ("Språkkommittén"), I wrote a program to do a query to get the titles for all degree project moments that had a project title. I then analyzed this data by year and by school.

1.1 Looking at LADOK data

Table 1 shows the number of grades for theses by school per year and also by cycle. This data is extracted from LADOK and only considers the degree project courses that were not canceled as of July 2021 (therefore, all early and canceled degree project courses were excluded). Figure 1 shows this data graphically. This data represents the number of degree projects for which a degree was entered in LADOK. For each entry, there is a title and an alternative title.

Table 2 shows the percentage of grades for which the title and the alternative title match (i.e., they are the same character string – which should be **unlikely** as one should be the title in Swedish and the other the title in English – and *vise versa*). Figure 2 shows this data graphically. With relatively rare exceptions (such as the title being the name of a place or the name of a project), the number of matching tiles and alternative titles should be rather low, as seen in the data for 1st cycle theses in CBH and ABE. However, even in these two cases, we see unexpected changes in the fraction of matching titles. This raised the question of whether this data is accurate. Table 3 examines this question for the LADOK data from EECS for the years 2019 and 2020.

There is also inconsistent use of "-" and ":" To separate the title from a subtitle in the LADOK titles.

Table 1: Number of grades for a project with a title – based upon data from LADOK as of 2021-08-09

School		Number of grades for a project with a title															
		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	total
1	ABE	0	8	29	259	254	277	344	376	374	303	339	465	446	470	300	4244
1	CBH	0	10	26	77	127	181	155	141	152	234	221	251	244	260	206	2285
1	EECS	0	0	5	252	301	318	325	361	369	382	421	443	478	496	455	4606
1	ITM	0	110	99	235	339	299	264	279	251	357	367	352	378	343	320	3993
1	SCI	0	88	105	159	151	183	196	222	245	259	223	271	227	187	210	2726
	total	0	216	264	982	1172	1258	1284	1379	1391	1535	1571	1782	1773	1756	1491	17854
2	ABE	7	120	209	245	376	411	411	374	359	361	395	439	378	437	221	4743
2	CBH	0	37	81	117	140	179	156	164	148	179	169	210	180	208	144	2112
2	EECS	2	79	109	129	127	172	135	120	170	397	512	568	588	669	281	4058
2	ITM	2	106	257	396	393	573	625	607	603	610	644	667	689	611	359	7142
2	SCI	0	38	89	111	120	168	188	223	250	246	260	330	313	320	189	2845
	total	11	380	745	998	1156	1503	1515	1488	1530	1793	1980	2214	2148	2245	1194	20900
	grand total	11	596	1009	1980	2328	2761	2799	2867	2921	3328	3551	3996	3921	4001	2685	38754

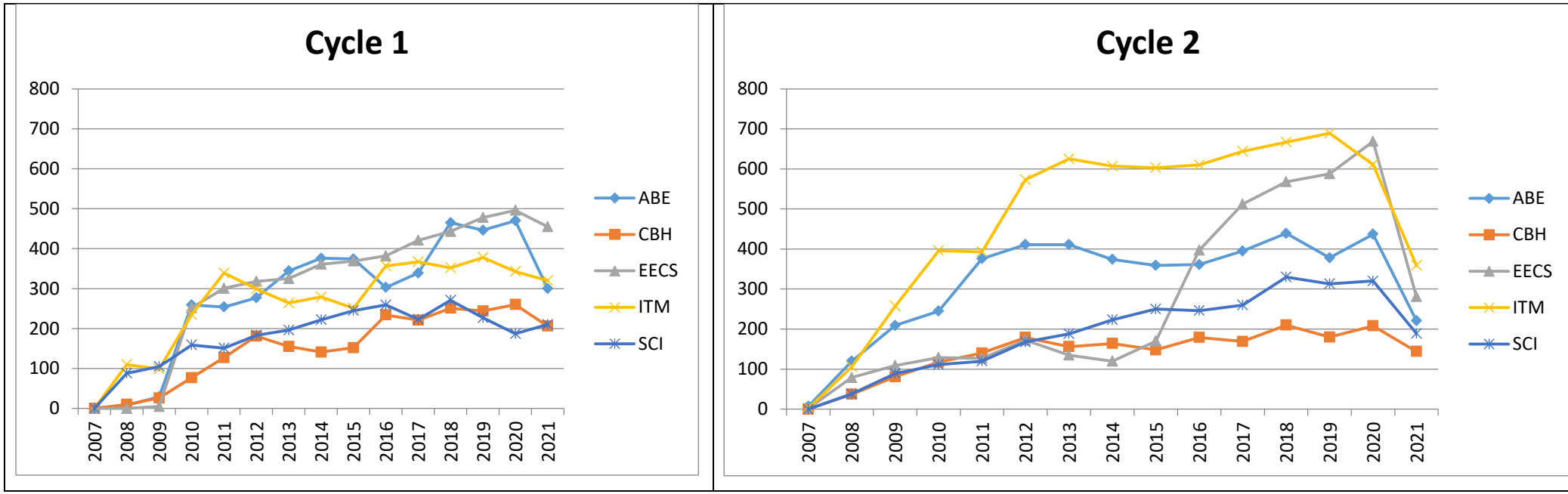

Figure 1: Number of grades for theses per year (first and second cycles)

Table 2: Fraction of matching title and alternative title in LADOK entries– based upon data from LADOK as of 2021-08-09

	School	Fraction of degree projects with titles that match																
		2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	median	last 5yr median
1	ABE			3.45	26.64	1.18	5.05	4.07	6.12	6.15	7.92	12.98	10.11	6.28	7.87	5.67	6.15	7.87
1	CBH			11.54	5.19	3.15	1.10	0.00	2.84	0.00	23.08	23.08	20.72	0.82	0.77	0.00	2.84	0.82
1	EECS			40.00	40.08	7.64	39.62	41.54	62.05	55.83	23.04	28.98	20.32	33.26	37.50	43.96	39.62	33.26
1	ITM		24.55	19.19	9.79	10.62	17.06	39.77	23.30	20.72	21.57	27.25	27.56	41.01	17.49	17.19	21.14	27.25
1	SCI		35.23	6.67	37.11	59.60	48.63	45.92	43.69	45.31	37.07	53.81	9.23	20.70	6.95	38.10	37.60	20.70
	median		29.89	11.54	26.64	7.64	17.06	39.77	23.30	20.72	23.04	27.25	20.32	20.70	7.87	17.19	20.71	
2	ABE	100.00	43.33	71.29	53.88	42.55	43.80	27.01	28.34	25.63	24.65	35.95	43.74	43.12	39.13	15.84	42.55	39.13
2	CBH		70.27	70.37	58.12	37.86	27.93	19.23	20.12	26.35	24.02	23.08	32.38	27.78	29.33	27.78	27.86	27.78
2	EECS	0.00	87.34	87.1	79.07	76.38	91.86	92.59	85.00	66.47	53.90	59.77	56.34	63.95	62.33	63.35	66.47	62.33
2	ITM	0.00	50.00	63.04	68.43	52.67	58.99	54.56	50.74	51.58	46.89	46.58	45.58	38.61	29.62	16.71	50.00	38.61
2	SCI		97.37	96.63	81.08	60.83	70.24	75.00	77.58	60.80	55.69	49.23	52.42	55.59	44.69	26.46	60.82	49.23
	median	0.00	70.27	71.29	68.43	52.6	58.99	54.56	50.7	51.58	46.89	46.58	45.58	43.12	39.13	26.46	50.74	

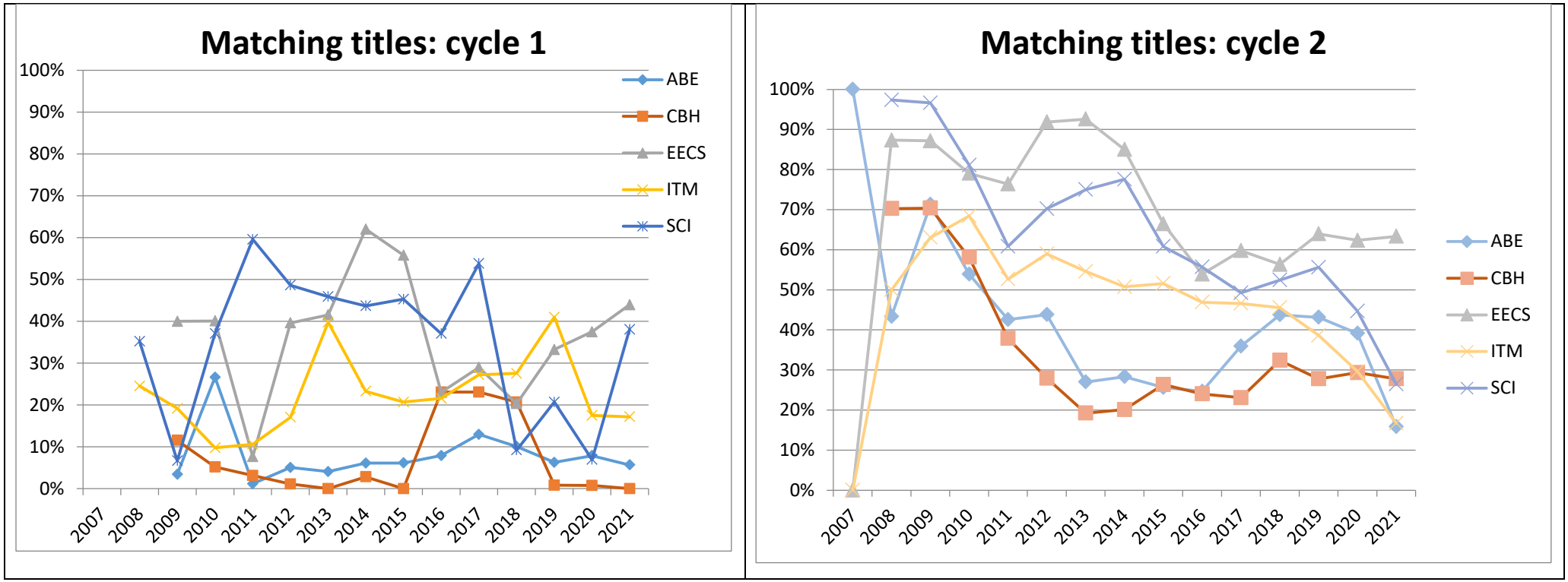

Figure 2: Fraction of matching title and alternative titles

1.2 Looking at DiVA data

The EECS numbers about matching titles from LADOK looked odd, as the number that did not have an English and a Swedish title had been going down until ~2015. So I looked closely at the results for 2020 using the data in DiVA and the full-text thesis from DiVA when it was available (I could not look at the 2021 data since many of the 2021 theses were not yet in DiVA). There are quite a lot of errors and omissions --- this data available on request.

Table 3 shows the results of checking whether the matching titles in LADOK match the data in DiVA. This examination of the theses with matching titles in LAODK considered the metadata in DiVA (during 2019) and the metadata in DiVA, and the actual thesis itself (2020). As we can see, in 2020, roughly 9.4% of the cases of the 601 theses with matching titles in LADOK do **not** reflect the data in DiVA or in the thesis itself. While in the case of 2019, roughly 2.2% of the theses with matching titles in LADOK do not have matching titles in DiVA's metadata. We can consider the case illustrated by 2019 data as representing a lower bound on the errors in titles, while the data from 2020 is probably more representative of the actual titles. Additionally, it should be noted that of the 600 theses with matching titles in LADOK in 2020, 11 of these theses do not have any record in DiVA (i.e., there is no entry for this thesis – not even the metadata)! Two theses claim to have full text, but I got an error about not being able to load the plugin when attempting to view the file. The full text was missing for an additional 14 theses. There are several theses have duplicates in DiVA:

as URN: urn:nbn:se:kth:diva-291546 and URN: urn:nbn:se:kth:diva-294066

as URN: urn:nbn:se:kth:diva-273926 and URN: urn:nbn:se:kth:diva-273266

as URN: urn:nbn:se:kth:diva-289523 and URN: urn:nbn:se:kth:diva-290794

One 1st cycle thesis has the correct Swedish title for one author, but not the other author! In one case, DiVA shows the English title again as the alternative title! In two cases, the inside cover page shows the Swedish title says: ": Detta är den svenska översättningen av titeln", i.e., the student used the new template but did not enter a Swedish title; hence they got the default text – and the examiner did not seem to have noticed! In one case, DiVA says the report is in Swedish, but it is in English and only has a title in English. In one case, the abstracts in English and Swedish are in the PDF but missing in DiVA.

In one case, the thesis has a Swedish sammananfaning before the abstract, although the text is in English. In one case, the abstract follows the TOC. In another case, the abstracts and keywords are numbered as sections and subsections. In some cases, the Swedish sammanfatning is missing.

Additionally, there are quite a number of cases where the cover is incorrect. Of these three have used a Swedish cover, but the thesis is actually in English. One thesis has both current and older covers! In a number of cases, it is clear that the cover is not correct with strings such as:

- "KTH Thesis Report" on the cover (3x),
- "KTH Master Thesis Report" (3x),
- "KTH Master Thesis" (1x),
- "Master Thesis Project" (2x)
- "KTH Bachelor Thesis Report" (2x), and
- "KTH Thesis Report" (2x).

In at least 4 cases, there is a back cover as the 2nd page in the PDF file. In one case, the Swedish keywords are actually English words. In 8 cases, there are spelling errors in the title.

Table 3: Does the DiVA data support the LADOK matching titles – The data for 2019 compares LADOK and DiVA data, while the data for 2020 extends the comparison to include the thesis itself (when the full-text was available via DiVA)

		Fraction of total theses in the year
2020		
460	accurate	0.375204
115	FALSE	0.093801
1226	total	
2019		
	accurate	Not examined
25	FALSE	0.022262
1123	total	

2 Could one exploit the DiVA data to check the quality of the LADOK data?

As we saw in the previous section, by looking at the DiVA metadata, it was possible to identify that some of the claimed matching titles were, in fact, not matching. This suggests that perhaps the DiVA metadata could be used to check the accuracy of the LADOK titles. However, this is **not** easy to do. The problem is that one needs an identifier available in both LADOK and DiVA, and there is no such identifier. Moreover, we cannot use the author's name as the author names are inconsistent between LADOK and DiVA.

However, it is possible to translate the LADOK identifier for a student to a KTH ID (hereafter kthid) – as this mapping is in Canvas. So if the DiVA entries have the kthid for the author or authors entered, then it would be possible to match the entries for the title entered with the grades in LADOK with the DiVA entry. However, as we can see in Table 4, the fraction of authors whose kthid was entered is too low to use this as a means to associate a LADOK title entry with a DiVA entry.

So it would appear that we do **not** have a simple means of connecting the two data sources unless:

- The DiVA administrators **enter the kthid for authors** and/or
- Some sort of fuzzy matching is done on author names and titles.

Table 4: Number of DiVA entries with KTHIDs for the authors

Some statistics for KTH from DiVA data about the number of student theses, with number of authors, and number of KTHIDs for these authors (data is taken from DiVA on 2021-07-25)

ABE theses with authors and KTHIDs

	2019	2020	2021
author0	646	725	382
author1	214	290	138
author2	2	1	1
with KTHIDs			
author0	8	4	20
author1	2	0	2
author2	0	0	0
percentage of authors with KTHIDs	1.16%	0.39%	4.22%

Totals students	2019	2020	2021
ABE	862	1016	521
CBH	225	376	156
EECS	1064	1186	146
ITM	991	900	392
SCI	510	512	220

CBH theses with authors and KTHIDs

	2019	2020	2021
author0	176	290	118
author1	49	68	37
author2	0	11	1
author3	0	7	0
with KTHIDs			
author0	7	12	1
author1	4	2	1
author2	0	0	0
author3	0	0	0
percentage of authors with KTHIDs	4.89%	3.79%	1.28%

KTHIDs	2019	2020	2021
ABE	10	4	22
CBH	11	14	2
EECS	21	7	0
ITM	14	3	0
SCI	3	2	10

EECS theses with authors and KTHIDs

	2019	2020	2021
author0	846	954	140
author1	218	230	6
author2		2	
with KTHIDs			
author0	15	4	0
author1	6	3	0
author2		0	
percentage of authors with KTHIDs	1.97%	0.59%	0.00%

ITM theses with authors and KTHIDs

	2019	2020	2021
author0	664	604	231
author1	320	290	152
author2	7	6	9
with KTHIDs			
author0	9	3	0
author1	4	0	0
author2	1	0	0
percentage of authors with KTHIDs	1.41%	0.33%	0.00%

SCI theses with authors and KTHIDs

	2019	2020	2021
author0	395	394	157
author1	114	118	63
author2	1	0	0
with KTHIDs			
author0	3	2	9
author1	0	0	1
author2	0	0	0
percentage of authors with KTHIDs	0.59%	0.39%	4.55%

3 How can you get the data and reproduce the results?

You can retrieve the DiVA data with commands of the form:

```
wget -O cbh-2019-diva.mods 'https://kth.diva-
portal.org/smash/export.jsf?format=mods&addFilename=true&aq=[[]]&aqe=[]&aq2
=[[{"dateIssued":{"from":"2019","to":"2019"}},{"organisationId":"879224","o
rganisationId-
Xtra":true},{"publicationTypeCode":["studentThesis"]}]]&onlyFullText=false&
noOfRows=5000&sortOrder=title_sort_asc&sortOrder2=title_sort_asc'
```

The organization IDs are:

'ABE': "5850",

'ITM': "6023",

'SCI': "6091",

'CBH': "879224", and

'EECS': "879223".

Followed by extracting the authors, titles, etc. using a command of the form:

```
./MODS_to_titles_and_subtitles.py --mods abe-2021-diva.mods
```

One can get the LADOK data for degree projects with a command of the form:

```
./thesis_titles_by_school.py -s EECS
```

All of the programs are available from https://github.com/gqmaguirejr/E-learning

To run the programs that get data from LADOK, you need the ladok3 python library. This library is available via https://pypi.org/project/ladok3/ with source code and examples of using it at https://github.com/dbosk/ladok3. The program also uses the KTH KOPPS API to get the data about the course codes and their status, and you may need an access key for this. Degree project course codes are assumed to be those that end in "X". The problem assumes that there is only a single course moment with a project title and alternative title and then uses this information as the thesis title and alternative title.