Deliverable Number: D8.4

Deliverable Title: Validation of ISCO-08 codes + explanatory note

Work Package: WP8

Deliverable type: Other

Dissemination status: Public

Authors:

Kea Tijdens, University of Amsterdam/AIAS

Casper Kaandorp, University of Amsterdam/AIAS

Date Submitted: April 2018

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.

Table of content

Executive summary

Current coding practices in multi-country surveys are mostly based on coding performed by a national survey agency and typically based on the national coding indexes. In most countries these indexes are prepared by the national statistical offices. For a multi-country validation of occupational coding activities of national survey agencies, the question arises whether the same occupational titles are coded into the same or in different ISCO-08 4-digit occupational units. For this SERISS deliverable we collected as many coding indexes as we could find, provided they used the ISCO-08 classification. The resulting merged validation database of coding indexes had 70,489 records of ISCO-08 5-digit occupational titles and their 4-digit code from 20 sources with 19 different languages. A check revealed that 10.3% of the records had codes that did not correspond with any codes in the official ISCO-08 classification.

To explore if similar occupational titles were coded similarly, all non-English occupational titles were translated into English using online dictionaries and Google translate. For 3,451 titles (4.9%) no translation was available. In the resulting validation database, 68% of the records were single titles and 32% of the titles had one or more duplicates, i.e. occurred at least twice in the combined database.

Our validation attempt had disappointing results. We expected that to a large degree the same occupational titles would be assigned the same code. Using the validation database, we applied two measures. First, of the 20,828 records with duplicates in more than one database or language, only 41% had the same code for the same occupational title. Second, we aggregated the records to occupational titles. Of the 5,754 titles slightly more than half (54%) had the same code for the same title. The remaining titles had not similar codes for similar occupational titles.

At the start of SERISS we had good hopes that we could add the database of national coding indexes to the cross-national occupation database developed for D8.3 and made available via www.surveycodings.org. However, the results of D8.4 indicate that this is not the right thing to do.

The merged validation database is available in the accompanying database ***SERISS-Deliverable 8-4 coding indexes 2018***.

1. Introducing SERISS

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) is a four-year project that aims to strengthen and harmonise social science research across Europe (2015-19). Work Package 8 (WP8) of SERISS aims to provide cross-country harmonised, fast, high-quality and cost-effective coding of open ended questions on respondents' occupations, industries and education into international standardized classification systems, and to develop a tool to collect standardized social network information. Occupation, industry, employment status, educational attainment and field of education are core variables in many socio-economic and health surveys. In addition, the size and intensity of social networks are key variables in social surveys. However, their measurement, especially in a cross-cultural, cross-national and longitudinal context, is cumbersome, not sufficiently standardized and often expensive. This work package takes recent scientific and technological developments as an opportunity to improve this situation for the benefit of survey measurement quality and to provide cost-effective solutions to Research Infrastructures (SERISS Annex 1, European Commission, 2015).

Building on the current technology and the partners' experiences, WP8 develops a cross-country harmonised, fast, high-quality and cost-effective coding module for the core variables, mentioned above. The module and its APIs use a large multi-lingual dictionary with tens of thousands of entries about job titles, industry names, fields of education and training, and employment status categories. Additionally, the module includes country-specific, structured lists of educational qualifications. The module provides up-to-date codes to classify the variables, using international standardized classification systems. It facilitates surveys in the ESS, EVS, GGP, SHARE and WageIndicator countries and their associated networks to serve infrastructures reaching out to a global audience. To support the ambition to strengthen the position of European infrastructures beyond the European Union, the occupational titles in the five most spoken languages outside the EU28 area have been included in the module, notably Russian, Mandarin, Arabic, Hindi and Bahasa. In total WP8 covers 47 languages servicing 99 countries. For the choice of countries and languages, see Deliverable D8.14 (Tijdens, 2016).

This report concerns Task 8.2 of WP8: "Compile the API-database of occupations". The responsible partner is the University of Amsterdam (UvA); partners are SHARE (UNIVE) and SHARE (CentERdata). The task aims to compile a database of occupational titles for 99 countries with in total 47 languages. All occupational titles are coded according to the ISCO-08 or International Standard Classifications of Occupations-08 (ILO, 2012). The database of occupational titles initially is derived from the WageIndicator occupation database with 1,700 multilingual occupational titles for 80 countries. Since 2000 the Netherlands WageIndicator web survey on work and wages uses a database of occupational titles for respondent's self-identification (Tijdens 2015). Gradually this database developed into a multilingual database, when from 2004 onwards more countries joined the survey. Translations were typically prepared by a national labour market expert in the national WageIndicator team. In 2008 the coding was adapted according to the update from ISCO-88 to ISCO-08. In 2015 the SERISS project facilitated the expansion of the database to 99 countries, to enlarge the number of occupational titles to approx. 4,000 and to provide the API to the research community on https://www.surveycodings.org/home .

Task 8.2 consists of five coherent deliverables:

- D8.3 Database of occupations + explanatory note
- D8.4 Validation of ISCO-08 codes + explanatory note
- D8.5 Vacancy crawler and additions to database + explanatory note
- D8.6 Job task collector and additions to database + explanatory note

- D8.7 Database of occupations for five languages + explanatory note

Deliverable D8.3 provides an occupational database with approx. 4,000 translated occupational titles into the 47 languages of the 99 countries selected for WP8. All titles are coded according to ISCO-08, in greater detail than 4-digit, which are hereafter referred to as 5-digit occupational titles. The source table has been generated from the ISCO-08 coding index (ILO, 2012). D8.4 takes another approach by merging coding indexes from many countries, validating their 4-digit codes across countries by translating the initial title into English. D8.5 relates to an attempt to use the occupation database to classify job titles from job vacancies into the ISCO-08 classification. D8.6 aims to analyse the coherence of tasks within ISCO-08 4-digits occupations as reported by survey respondents in the WageIndicator web survey. The work for D8.7 was outsourced to the Institute for Employment Research (IER), University of Warwick, UK, to provide the translations of the occupation database for Russian, Mandarin, Arabic, Hindi and Bahasa, and a validation check of the coding.

Whereas D8.3 relies on translations of 5-digit occupational titles from English into the non-English languages to build a cross-national database, D8.4 follows a reverse approach. In this Deliverable a database is compiled of 5-digit occupational titles of ISCO-08 coding indexes available from National Statistical Offices from predominantly non-English speaking countries. For D8.4 the occupational titles in these indexes have been translated into English, and their ISCO-08 codes have been validated by comparing these English titles. In this paper we refer to this database as ‘merged validation database D8.4’.

The responsible partner of Deliverable D8.4 is the University of Amsterdam/AIAS (UvA). Two persons from UvA made up a team to prepare and conduct the validation, notably Kea Tijdens, researcher, and Casper Kaandorp, software programmer.

2. Introducing deliverable D8.4

Current coding practices

Current practices regarding the occupation question in multi-country surveys mostly apply to coding performed by a national survey agency. These coding practices are typically based on the national coding indexes. Occupational titles as reported by survey respondents are classified in ISCO-08 occupational units. Note that ISCO-08 is a four-level hierarchical classification with ten major groups at the top of the hierarchy (1-digits), and 436 occupational units at the bottom, the so-called 4-digit units. Survey respondents can report up to ten thousands of occupational titles, which are to be classified into ISCO-08 4-digit.

Our study aims for occupational titles beyond the 4-digit classification, thus 5-digit occupational titles. Survey respondents reply to the question 'what is your occupation?' in 5-digit occupational titles, which are close to their actual job titles. They never reply in 4-digit categories, because these are too highly aggregated from their job titles. Survey agencies therefore face the challenging task to code 5-digit titles into a 4-digit classification. For this reason, it is important to explore the mapping from 5-digit to 4-digit classifications.

In most countries the national coding indexes are prepared by the national statistical office (NSO). National coding indexes usually group a set of occupational titles into one ISCO-08 4-digit unit group. Typically, these NSO's have to manoeuvre between previous and current classifications to ensure comparability over time and between the ISCO-08 classification and its proposed detailed list of occupational titles (ILO, 2012). For a multi-country comparison of the occupational coding activities of national survey agencies, the question arises whether the same 5-digit occupational titles are coded into the same ISCO-08 4-digit occupational units. Note that ILO does not provide a multilingual coding index, only an English index. Occupational coding in multi-country surveys is therefore basically a black box, with no answer given to the question whether the same occupations are coded similarly across countries.

Two viewpoints exist when comparing occupational classifications across countries and languages. The first one states that occupational titles cannot be translated beyond ISCO-08 4-digit, because the design of occupational boundaries follows country-specific practices and so does the composition of the country's industry. Therefore, the classification of thousands of 5-digit occupational titles into the ISCO-08 4-digit classification cannot be validated by means of a cross-country comparison. Hence, validating the occupational coding in multi-country surveys is not possible. The second viewpoint states that occupational titles can be translated, allowing for a cross-country validation of the occupational coding. This viewpoint assumes that across countries similar occupations exists. An argument in favour of the latter viewpoint assumes a convergence across countries that could be explained by factors such as the fast increase in global equipment suppliers which causes that equipment-defined tasks in jobs become similar across countries, the globalisation of the economy and the related need to understand occupations across countries, and the pressure towards cross-country standardization, e.g. QESH auditor. So far, very few studies, if any at all, have explored which viewpoint is true, though it may well be the case both viewpoints can be true at the same time for parts of the occupational titles, e.g. depending on industries.

A cross-country comparison of occupational titles is typically done by translating the initial titles into English, and to compare these translations. This deliverable, D8.4, aims to validate whether occupational titles listed in national coding indexes are coded similarly across countries. For this purpose, coding indexes from many countries have been merged, the initial titles have been translated in English, and their 4-digit ISCO-08 codes have been compared across these countries. This paper explains the methods applied and their results.

Validating multi-country coding practices relies on:

a) the availability of coding indexes of NSO's, particularly for non-English speaking countries
b) the detail of these indexes: do they specify 5-digit occupational titles within each 4-digit unit group
c) the availability of translations of these 5-digit occupational titles into English and the quality of these translations
d) a text mining program that identifies whether the translated English occupational titles are the same, e.g. *office clerk* equals *clerk in the office*.

Research objectives

Two objectives are central for the validation exercise in deliverable D8.4:

1) To what extent can translated occupational titles be compared?
2) What percentage of similar occupational titles is coded similarly across countries?

The activities undertaken for the validation consist of several steps, which are discussed in the next section.

3. Validating occupational coding across countries

Step 1 Collecting coding indexes

In a first step, we collected as many coding indexes we could find, as long as these came from countries included in the list of 99 countries, as defined for the SERISS project (see Table 6-1 in Appendix). For these 99 countries we collected coding indexes from the National Statistical Offices (NSO). Starting from the UN's website with lists of national classifications for many countries, we explored the websites of the NSO's to find their coding indexes. For a few countries, where no index was available online, we were able to retrieve a coding index from researchers in that country. Altogether we found coding indexes for 34 countries (Table 6-1 in Appendix). Of these 34 countries, five had another classification than ISCO-08 (Canada, Iceland, India, Italy, Switzerland). These countries either used ISCO-88 or had their own national coding index, e.g. Canada. Four countries had no coding index beyond the 4-digits of ISCO-08 (Croatia, Malaysia, Norway, Tanzania) and could therefore not be used. Two countries referred to other countries regarding their ISCO-08 coding indexes. Germany referred to the Austrian coding index and Montenegro stated that its index also applies to Serbia. For four countries we encountered technical difficulties. The coding index of Greece was incomplete, as it did not go beyond ISCO-08 code 2122. Translating Hebrew caused too many technical difficulties, among others due to the right-to-left script. Israel therefore could not be included. Due to a technical error the Belgium and the Lithuanian coding indexes were dropped from our validation database.

Out of the initial 34 coding indexes, 19 could be merged into the validation database. One country, Finland, had an index in two languages (Finnish and Swedish). Because we already had a long Swedish occupational coding index from Statistics Sweden, we did not include the brief Swedish occupational list from Finland. Finally, we added the English occupational titles of the WageIndicator/SERISS database (the occupational database from D8.3, Tijdens, 2019, forthcoming). The resulting merged database of coding indexes had 70,489 records of occupational titles from 20 sources with 19 different languages (see Table 3.1).

As Table 3.1 shows, the number of occupational titles in the validation database varies largely across the source countries. Austria contributes more than 13,000 occupational titles, which comprises 19% of the total database. Sweden ranks second with more than 12,000 titles (12%). Albania, Bulgaria, Estonia, Latvia, Netherlands, WageIndicator contribute each 4,000 to 5,000 titles (6-7%). Czech Republic, Montenegro, Poland, Romania, Slovak Republic, Slovenia, Spain, South Africa, Turkey contribute each 1,000 to 2,500 titles. The coding indexes of three countries (Denmark, Finland, Portugal) comprise less than 1,000 titles.

Table 3-1 Columns in the merged validation database of translated occupational titles

#	Database id	Coding index/source	Language	# entries in index	%
1	AL_org	Albania	Albanian	4273	6.1
2	AT_org	Austria (+ Germany)	German	13395	19.0
3	BG_org	Bulgaria	Bulgarian	5077	7.2
4	CZ_org	Czech Republic	Czech	1358	1.9
5	DK_org	Denmark	Danish	564	0.8
6	EE_org	Estonia	Estonian	4715	6.7
7	FI_org	Finland	Finnish	103	0.1
8	LV_org	Latvia	Latvian	4057	5.8
9	ME_org	Montenegro (+ Serbia)	Serbian	2129	3.0
10	NL_org	Netherlands	Dutch	4704	6.7
11	PL_org	Poland	Polish	2443	3.5
12	PT_org	Portugal	Portuguese	708	1.0

#	Database id	Coding index/source	Language	# entries in index	%
13	RO_org	Romania	Romanian	3077	4.4
14	SK_org	Slovak Republic	Slovak	2147	3.0
15	SI_org	Slovenia	Slovenian	2094	3.0
16	ES_org	Spain	Spanish	2502	3.5
17	SE_org	Sweden	Swedish	8617	12.2
18	ZA_org	South Africa	English	2252	3.2
19	TR_org	Turkey	Turkish	2214	3.1
20	ENG_org	WageIndicator	English	4060	5.8
	TOTAL			70489	100%

Step 2 Non-existent codes

After the occupational titles and their ISCO-08 4-digit codes from the 20 indexes were merged, we checked the existence of the codes by comparing these to the codes available in the official ISCO-08 coding index (ILO, 2012). It turns out that 10.3% of the 70,489 titles had codes that do not correspond with any codes in the ISCO classification. We call them non-existent codes. Only in nine of the twenty indexes did we not encounter any non-existent codes, notably Austria, Bulgaria, Czech Republic, Estonia, Finland, Latvia, the Netherlands, Slovak Republic, and Turkey (Table 3.2). In five countries we find a few, notably in Denmark, Montenegro, Slovenia, South-Africa, and WageIndicator. In six countries we noticed substantial numbers of non-existent codes. In the Spanish coding index, 69% of the occupational titles had such codes, followed by Romania with 59%. Sweden had 29%, Albania 24%, Portugal less than 8%, and Poland around 6%. We did not ask the NSO's for the reasons of using non-existent codes, but most likely the non-existent codes are used to enable mapping with their long-existing national classification, which was not harmonized with the ISCO classification.

Table 3-2 Mean score for percentage existent code by country

country	Mean existent code	N	Std. Dev
Albania	75.8%	4273	0.42834
Austria, ref. Germany	100.0%	13395	0
Bulgaria	100.0%	5077	0
Czech Republic	100.0%	1358	0
Denmark	97.9%	564	0.14443
Estonia	100.0%	4715	0
Finland	100.0%	103	0
Latvia	100.0%	4057	0
Montenegro, ref. Serbia	100.0%	2129	0.02167
Netherlands	100.0%	4704	0
Poland	94.1%	2443	0.23633
Portugal	92.2%	708	0.26786
Romania	40.9%	3077	0.4917
Slovak Republic	100.0%	2147	0
Slovenia	98.3%	2094	0.13002
Spain	31.4%	2502	0.46427
Sweden	72.1%	8617	0.44835
South Africa	97.6%	2252	0.15439
Turkey	100.0%	2214	0
WageIndicator	100.0%	4060	0.01569
Total	89.7%	70489	0.30423

A breakdown by the ten ISC0-08 1-digit groups, the so-called major groups, provides details for these results (Table 3.3). Explanations follow after the table.

Table 3-3 Mean score for percentage existent code by ISCO-08 1 digit major group

ISCO-08	Label 1dgt major group	Mean existent code	N	Std Dev
0	Armed forces occupations	0.5698	516	0.49559
1	Managers	0.8182	5237	0.38570
2	Professionals	0.8916	16306	0.31085
3	Technicians and associate professionals	0.9370	13108	0.24300
4	Clerical support workers	0.8470	3039	0.36006
5	Services and sales workers	0.8981	4680	0.30258
6	Skilled agricultural, forestry and fishery workers	0.9188	2106	0.27320
7	Craft and related trades workers	0.9277	12906	0.25898
8	Plant and machine operators and assemblers	0.8813	9437	0.32343
9	Elementary occupations	0.8916	3154	0.31098
Total		0.8968	70489	0.30423

Major group 0, the *Armed forces occupations*, have the lowest percentage of existent codes, notably 57%. This is due to three countries. In Albania and Portugal all army occupations are coded to a greater extent than in the ISCO-08 classification, and in Spain only 76% are coded with existent codes. Take note that Finland, Romania and South-Africa have no occupational titles for this ISCO-08 major group 0.

Major group 1, the *Managers*, has the second lowest percentage of existent codes, notably 82%. This is due to five countries. The coding index of Romania reveals only 16% existent codes, followed by Sweden with 35%, and Albania with 71%. The remaining two countries with non-existent codes are South-Africa (92% existent codes) and Slovenia (96%).

Major group 2, the *Professionals*, has 89% occupational titles with an existent code, which is due to non-existent codes in seven countries. The country with the lowest percentage of existent codes is Romania (42%), followed by Sweden (71%), Albania (82%), and Poland (89%). The remaining three countries, Denmark, Slovenia and South Africa, have rates above 90%.

Major group 3, the *Technicians and associate professionals*, has 94% occupational titles with an existent code. Six countries have non-existent codes. The country with the lowest percentage of existent codes is Romania (59%), followed by Albania (81%), and Sweden (88%). The remaining three countries, Poland, Slovenia and WageIndicator, have rates above 90%.

Major group 4 is the group of the *Clerical support workers*. Only three countries have non-existent codes. The country with the lowest percentage of existent codes is Romania (35%), followed by Albania (38%), and Sweden (55%).

Major group 5 concerns the *Services and sales workers*. Four countries have non-existent codes. The country with the lowest percentage of existent codes is Romania (53%), followed by Albania (75%), Sweden (74%), and South-Africa (98%).

Major group 6 is the group of the *Skilled agricultural, forestry and fishery workers*. Three countries have non-existent codes. The country with the lowest percentage of existent codes is Romania (44%), followed by Albania (70%), and South-Africa (99%).

Major group 7 concerns the *Craft and related trades workers*. Seven countries have non-existent codes. The country with the lowest percentage of existent codes is now Spain

(34%), followed by Romania (44%), Albania (73%), and Sweden (88%). The remaining countries Montenegro, Poland, Slovenia all have rates above 98%.

Major group 8 is the group of *Plant and machine operators and assemblers*. Six countries have non-existent codes. The country with the lowest percentage of existent codes is again Romania (34%), followed by Spain (59%), and Sweden (64%). South Africa and Albania both have 90%, and Poland has 98%.

Major group 9 is the group of *Elementary workers*. Four countries have non-existent codes. Spain has a very low percentage of existent codes (5%). Romania has 27%, followed by Sweden (59%) and South Africa (98%).

In conclusion, the *Armed forces occupations* are a problematic major group, because the coding indexes of some countries do not provide any detailed classification in this group, whereas some other provide a classification with codes, not defined in ISCO-08.

Step 3 Translating the occupational titles

The objective of our validation exercise is to explore whether similar 5-digit occupational titles in different countries are coded similarly. Similarity is here defined as a similar English translation of the titles in the national coding indexes. For the concept of 'similar occupations' we fully rely on the occupational title as any information concerning the job content is absent.

To explore if similar occupations are coded similarly, all non-English occupational titles from the coding indexes have been translated into English. Given the huge number of 70,489 occupational titles, translation by professional translators was beyond our budget. Instead, we developed scripts for using online dictionaries and, where these dictionaries did not provide a translation, Google translate. The database could be extended with a column with English translations. Two coding indexes had already English occupational titles: South Africa and WageIndicator (jointly 6,312 titles, or 9%).

The merged validation database has 70,489 records of occupational titles. For 3,451 titles (4.9%) the online dictionaries and Google translate do not provide a translation. Three countries have relatively high percentages of non-translatable occupational titles: Austria (12.2%), the Netherlands (7.4%), and Sweden (9.9%). The more titles in a national coding index, the higher the percentage of non-translatable titles (r=.80). For example, the 13,395 occupational titles in the Austrian index are more likely to include some country-specific, and therefore untranslatable, titles than the 103 titles in the Finnish index. Translatability does not relate to the number of characters within a single occupational title.

Table 3.4 provides an example of the translations. Six national coding indexes have 5-digit occupational titles that have been translated into the English occupational title *art historian*. The first column of the table shows that three of the six coding indexes assign 4-digit code 2633 *Philosophers, historians and political scientists* to this occupational title, whereas two others assign code 2651 *Visual artists* and one 2632 *Sociologists, anthropologists and related professionals*. This coding dissimilarity will be discussed in Section 4 of this paper.

Table 3-4 Example of the translation of the national occupational titles into English titles

isco-08 4 digit	coding index from	national label	English translation
2632	11 'Netherlands'	kunsthistoricus	art historian
2633	12 'Poland'	historyk sztuki	art historian
2633	2 'Austria (de)'	kunsthistoriker	art historian
2633	9 'Latvia'	mākslas vēsturnieks	art historian
2651	6 'Estonia'	kunstiteadlane	art historian
2651	16 'Slovenia'	umetnostni zgodovinar	art historian

Step 4 Preparing the validation database

For the similarity test we exclude the non-translated occupational titles and those with non-existent codes. We therefore can use 70,489 - 7,275 - 2,655 = 60,559 records (see table 3.5).

Table 3-5 Number of records in the database, broken down by translation and existence

	Non-existent	Existent	Total
Original English	56	6256	6312
Not translatable	303	2655	2958
Translated	6916	54303	61219
Total	7275	63214	70489

To improve the similarity matches of occupational titles, we apply a three-step script-based approach to clean the English translations:

a) removing dots, non-alphabetic characters, double blanks, quotes, and redundant words as the, of the, an, and alike,
b) changing occupational titles from plural to singular and from female to male titles,
c) moving words to other parts of the sentence if that provided a similarity match,
d) improving the identification of similar titles by creating a second column with the English occupational titles in which all blanks are removed.

Records with English titles that occur at least twice in the database are assigned a similarity score of 1 whereas the records without duplicates are assigned value 0. The results will be discussed in the next section.

The validation database can be found in the accompanying database (in SPSS and excel) of this Deliverable (*SERISS-Deliverable 8-4 coding indexes 2018).* This database has 70,489 records and has nineteen variables (see Table 3.6).

Table 3-6 Overview of variable names and labels in accompanying database (SERISS-Deliverable 8-4 coding indexes 2018).

Variable name	Variable label
ISCO08	ISCO-08 4_dgt code from the coding indexes
ISCO08_1dgt	ISCO-08 1_dgt code derived from 4_dgt code
country	Source of the coding index
label_NATLANG	Occupational title in national language
label_Engels	Occupational title translated in English
label_Engels_NOBLANKS	Occupational title translated in English_without blanks
length_NATLANG	Number of characters in national title from coding index
duplicate_label_eng	Value 1 if at least twice present in the English translations, value 0 if not so
existingISCO084dgt	Value 1 if code from coding index is existing ISCO-08 code, value 0 if not so
translated	Value 1 if translated from national language into English, value 0 if title was not translatable
similarity	Value 1 if all duplicate titles have same ISCO-08 code
Mana	Managers
Prof	Professionals
Tech	Technicians and associate professionals
Cler	Clerical support workers
Serv	Service and sales workers
Agri	Skilled agricultural, forestry and fishery workers
Craft	Craft and related trades workers
Machi	Plant and machine operators and assemblers

4. Results

To what extent can translated occupational titles be compared?

Our first objective addresses the extent to which translated occupational titles can be compared. As said, of the merged database of 70,489 records, we use 60,559 records to explore this objective. As table 4.1 shows, 32% of the records have titles that are at least twice present in the database, and 68% are single titles. For reasons of readability we use the words *duplicate titles* for the records with titles that are at least twice in the database. The 68% non-matching occupational titles is much higher than expected. Several factors may explain why this percentage is so high.

Table 4-1 Frequency and percentage of single and duplicate occupational titles

duplicate_label_eng	Frequency	Percent
Single title	41369	68.3
Duplicate title	19190	31.7
Total records	60559	100.0

The first explanation is that the percentage of duplicate translations varies largely across countries, ranging from 4% of the English translations of the Danish titles to 68% of the translations of the Turkish occupational titles and 84% of the South-African English titles (see Table 4.2). The percentage of duplicates is not related to the number of entries in the coding index (r=.08). We concluded that some countries adhere to the ISCO-08 coding index to a much larger extent than other countries do, which hampers the similarity tests.

Table 4-2 Percentage titles with at least one duplicate title

country	Mean duplicate	N	Std. Dev.
1 Albania	.16	3163	.366
2 Austria (de)	.33	11946	.469
3 Bulgaria	.16	5077	.369
4 Czech Republic	.11	1358	.308
5 Denmark	.04	540	.202
6 Estonia	.31	4702	.464
7 Finland	.20	103	.405
9 Latvia	.25	3998	.434
10 Montenegro (sr)	.38	2085	.485
11 Netherlands	.31	4481	.464
12 Poland	.23	2213	.422
13 Portugal	.27	651	.445
14 Romania	.16	1222	.370
15 Slovak Republic	.16	2126	.368
16 Slovenia	.35	2036	.476
17 Spain	.38	783	.485
18 Sweden	.28	5609	.451
19 South Africa (en)	.84	2197	.371
20 Turkey	.68	2210	.468
21 WageIndicator (en)	.48	4059	.499
Total	.32	60559	.465

A second explanation relates to the fact that some languages use brief sentences instead of one or two words to identify an occupational title. For example, the Bulgarian occupation *Работник, ремонт на стоматологични инструменти* is translated into *Worker repairing dental instruments.* In the same way the Serbian title *Radnik na proizvodnji kompota* is translated as *A worker on the production of compote.* These sentence-like occupational titles hamper the similarity test. On average, the initial occupational titles count 25.2 characters. The titles for which a duplicate is identified count on average 17.5

characters, whereas the titles without a duplicate count 28.9 characters (Table 4.3). The table reveals that the occupational titles in the Czech, Danish, Slovak and Portuguese indexes have the largest number of characters. Testing the similarity of occupational titles is thus hampered by different name-giving concepts of occupational titles across countries.

Table 4-3 Mean number of characters of the national occupational titles (length)

	Duplicate		Total	N	Std. Dev.
Mean length	no	Yes			
1 Albania	32.6	16.3	30.0	3163	14.0
2 Austria (de)	23.0	17.7	21.3	11946	11.2
3 Bulgaria	33.2	17.4	30.6	5077	16.6
4 Czech Republic	42.5	19.1	40.1	1358	20.6
5 Denmark	40.1	24.6	39.5	540	17.4
6 Estonia	22.2	15.7	20.1	4702	8.8
7 Finland	30.9	20.8	28.8	103	12.5
9 Latvia	31.9	17.7	28.3	3998	17.3
10 Montenegro (sr)	28.8	17.1	24.4	2085	12.1
11 Netherlands	24.7	15.4	21.8	4481	10.4
12 Poland	34.8	15.4	30.3	2213	15.6
13 Portugal	45.1	18.5	37.9	651	21.0
14 Romania	32.2	16.2	29.6	1222	16.0
15 Slovak Republic	43.1	16.6	38.9	2126	21.5
16 Slovenia	35.0	18.1	29.1	2036	18.1
17 Spain	39.9	24.8	34.2	783	16.9
18 Sweden	21.9	15.9	20.2	5609	9.6
19 South Africa (en)	23.1	18.9	19.6	2197	8.7
20 Turkey	26.4	20.1	22.1	2210	11.5
21 WageIndicator (en)	25.9	18.0	22.2	4059	11.0
Mean	28.9	17.5	25.2		
N	41369	19190		60559	
Std. Deviation	15.5	8.8			14.7

A detailed insight into the likelihood of an occupational title to have a duplicate is shown in Table 4.4. We assume that the number of characters in the initial occupational title decreases the likelihood of a duplicate title. Indeed, the regression results show that this effect holds, also when controlled for ISCO08 1-digit groups.

Table 4-4 Logistic regression for the binary variable duplicate (1=duplicate present, 0 = no duplicate present) by length of initial wording and by ISCO-08 1-digit categories (Ref: armed forces, elementary and agricultural occupation), N= 60,559

Variables	Exp(B)	Sig.	Exp(B)	Sig.
Length: number of characters in national title	0.912	***	0.908	***
Managers			2.050	***
Professionals			2.007	***
Technicians and associate professionals			1.600	***
Clerical support workers			1.114	
Service and sales workers			1.236	***
Craft and related trades workers			0.804	***
Plant and machine operators and assemblers			0.850	***
Constant	3.526	***	2.889	***

A third explanation relates to the large variation of occupational titles across coding indexes per ISCO-08 4-digit unit. Within occupational unit 9123 *Window cleaners* the twenty countries provide between zero and two 5-digit occupational titles, and as a consequence the number of duplicate translations is high in this 4-digit occupational unit. In contrast, within occupational unit 8160 *Food and related products machine operators* the twenty countries provide between 0 and 133 5-digit occupational titles, ranking Austria with 133 and Spain with 122 titles on top. With 211 titles in 4-digit occupational unit 5223 *Shop sales assistants*

Austria has the maximum number of 5-digit occupational titles within a single ISCO-08 4-digit unit. The chance of finding similar English translations is therefore much smaller in this unit.

In conclusion, in the validation database of 60,559 records, for a high percentage of 68% titles no similar occupational title could be identified. One of the reasons is the large variation across countries and across ISCO-08 4-digit occupational units in the degree to which these units are specified. In addition, a higher number of characters in the initial occupational title decreases the likelihood of a duplicate title.

Similar occupational title, not similar code

The second objective of our study is to explore what percentage of similar occupational titles is coded similarly across countries, thus have been assigned similar ISCO-08 4-digit codes in the national coding indexes. We develop a similarity index for the occupations that are at least twice present in the database (duplicate = yes). This part of the validation exercise is limited to the 19,190 records with duplicate occupational titles in the database. We also included the 1,638 records with non-existent codes but with duplicate titles, in total 20,828 records.

Let's start first with an example. Table 4.5 shows the classification problems for a selection of five 5-digit occupations in the database. Three coding indexes include the occupation *wine consultant*, two of which are assigned code 5131, and one code 7613. The similarity index for the *wine consultant* is therefore 67%. Eight coding indexes include the occupation *wine grower*, of which six are assigned code 6112 and the other two code 6110 and 6111, resulting in a similarity of 75%. Six coding indexes include the occupation *wine maker*, of which two are assigned code 6112, and the others the codes 2145, 7514, 7515, and 8169, resulting in a similarity of 33%. Four coding indexes include the occupation *wine waiter*, of which three are assigned code 5131, and the other code 5181, resulting in a similarity of 75%. According to Table 4.5, for five English occupational titles 16 different codes in the national coding indexes are notified for a total of 30 national occupational titles, resulting in similarities ranging from 33% to 75%.

Table 4-5 Selection of five occupational titles and their ISCO-08 codes

ISCO code	label	Count	total	max	similarity
5131	Wine consultant	2	3	2	67%
7613	Wine consultant	1			
6112	Wine grower	6	8	6	75%
6110	Wine grower	1			
6111	Wine grower	1			
6112	Wine maker	2	6	2	33%
2145	Wine maker	1			
7514	Wine maker	1			
7515	Wine maker	1			
8169	Wine maker	1			
7515	Wine taster	5	9	5	56%
7709	Wine taster	2			
7517	Wine taster	1			
7613	Wine taster	1			
5131	Wine waiter	3	4	3	75%
5181	Wine waiter	1			
TOTALS	5 English occupations	30 national occ			

The final piece of our validation exercise of occupational coding across countries presents the overall results of the dissimilarity testing (Table 4.6). The merged database has 20,828 records with occupational titles that were present at least twice in the database. Fewer than

half of these records have duplicates which have 100% the same occupational title, resulting in a similarity of 41%.

Table 4.6 shows further that the 20,828 records could be aggregated into 5,754 occupational titles, hence 3.6 records per title. Of these titles 3,131 have a 100% similar code, applying to slightly more than half of the titles (54%). Another 1,183 titles have a similar coding for 51-99% of the titles (21%). And another 1,132 titles have a 50% similar coding (20%). Finally, 308 occupational titles have several codes for the same occupational title.

Table 4-6 Number of duplicate records and number of duplicate occupational titles

Records	#
Records present at least twice, incl records with non-existent codes	20828
Of which the duplicates had for 100% the same code	8619
Of which the duplicates had for less than 100% the same code	12209
Similarity across the 20828 records	41%
Titles	#
Number of English occupational titles covered by the 20828 records	5754
Of which all duplicates were coded into one title (100% similar)	3131
Of which duplicates were coded similar: 51-99%	1183
Of which duplicates were coded similar: 50%	1132
Of which duplicates were coded similar: <50%	308
Average similarity across the 5754 titles	81%

5. Conclusions

In conclusion, our validation attempt leads to disappointing results. We expected that to a large degree the same occupational titles would be assigned the same code in the national coding indexes. Using the validation database, we applied two measures, to test whether this was indeed the case. First, of the 20,828 records with duplicates, only 41% had the same code for the same occupational title. Second, we aggregated the records to occupational titles. Of the 5,754 titles slightly more than half (54%) had the same code for the same title. The remaining titles did not have similar codes for similar occupational titles.

At the start of SERISS we had good hopes that we could add the database of coding indexes to the cross-national occupation database developed for D8.3. However, the results of D8.4 indicate that this is not the right thing to do. Therefore, in the database we will include only the occupational titles translated similarly in at least two different languages and coded similarly.

For multi-country surveys that rely on open-ended surveys questions about occupation and coding by national survey agencies our study indicates that analyses rather should use the 3-digit ISCO classification than the 4-digit one. The limitations of databases coded at national level for comparative study also serves to emphasise the value of the work being carried out under the SERISS project to build a harmonised cross-national database of occupations.

6. References

CBS (2013) Codelijsten en beroepenindex ISCO 2008. Voorburg, Centraal Bureau voor de Statistiek

European Commission, Directorate-General for Research & Innovation, Research infrastructure (2015) ANNEX 1 (part A) Research and Innovation action NUMBER — 654221 — SERISS, Brussels

ILO (2012) International Standard Classification of Occupations ISCO-08. Volume 1 Structure, Group Definitions and Correspondence Tables. Geneva: International Labour Office

Tijdens KG (2015) Self-identification of occupation in web surveys: requirements for search trees and look-up tables, Survey Methods: Insights from the Field, DOI:10.13094/SMIF-2015-00008

Tijdens KG (2016) Survey Q&A + explanatory note. Deliverable D8.14 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables

Tijdens KG (2017) Database of occupations for five languages + explanatory note. Deliverable D8.7 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables

7. Appendix

Table 7-1 Overview of 99 countries, the availability of a coding index, the coding index for ISCO-08, ISCO-08 5 digit, no duplicate of other countries index, no technical difficulties, number of entries available at 5 digit ISCO-08

COUNTRY	locale	coding in dex avail.	ISCO-08	ISCO-08 5 dgt	no du- plicate	no tech - ical diff.	# entries 5 dgts ISCO-08
Albania	sq_AL	1	1	1	1	1	4273
Angola	pt_AO	0					
Argentina	es_AR	0					
Australia	en_AU	0					
Austria	de_AT	1	1	1	1	1	13395
Belarus	be_BY	0					
Belgium	nl_BE/fr_BE	1	1	1	1	0	
Benin	fr_BJ	0					
Bolivia	es_BO	0					
Bosnia and Herzegovina	bs_BA	0					
Brazil	pt_BR	0					
Bulgaria	bu_BU	1	1	1	1	1	5077
Burundi	fr_BI	0					
Cambodia	km_KH	0					
Cameroon	fr_CM	0					
Canada	en_CA	1	0				
Chile	es_CL	0					
China	zh_CN	0					
Colombia	es_CO	0					
Costa Rica	es_CR	0					
Croatia	hr_HR	1	1	0			
Cyprus	tr_CY	0					
Czech Rep.	cs_CZ	1	1	1	1	1	1358
Denmark	da_DK	1	1	1	1	1	564
Ecuador	es_EC	0					
Egypt	ar_EG	0					
El Salvador	es_SV	0					
Estonia	et_EE	1	1	1	1	1	4715
Ethiopia	am_ET	0					
Finland	fi_FI/sv_FI	1	1	1	1	1	103
France	fr_FR	0					
Germany	de_DE	1	1	1	0		
Ghana	en_GH	0					
Greece	gr_GR	1	1	1	1	0	
Guatemala	es_GT	0					
Guinea	fr_GN	0					
Honduras	es_HN	0					
Hungary	hu_HU	0					
Iceland	is_IS	1	0				
India	hi_IN	1	0				
Indonesia	ba_ID	0					
Iran	fa_IR	0					
Ireland	en_IE	0					
Israel	he_IL	1	1	1	1	0	
Italy	it_IT	1	0				
Japan	jp_JP	0					
Kazakhstan	ru_KZ	0					
Kenya	en_KE	0					
Kosovo	sq_KO	0					
Latvia	ar_IQ	1	1	1	1	1	4057
Lithuania	lt_LT	1	1	1	1	0	
Luxembourg	de_LU	0					
Macedonia	mk_MK	1	1	1	0		
Madagascar	fr_MG	0					
Malawi	en_MW	0					

COUNTRY	locale	coding in dex avail.	ISCO-08	ISCO-08 5 dgt	no du- plicate	no tech - ical diff.	# entries 5 dgts ISCO-08
Malaysia	ms_MY	1	1	0			
Malta	en_MT	0					
Mexico	es_MX	0					
Moldova	ro_MD	0					
Montenegro	sr_ME	1	1	1	1	1	2129
Mozambique	pt_MZ	0					
Netherlands	nl_NL	1	1	1	1	1	4704
New Zealand	en_NZ	0					
Nicaragua	es_NI	0					
Nigeria	en_NG	0					
Norway	no_NO	1	1	0			
Pakistan	en_PK	0					
Paraguay	es_PY	0					
Peru	es_PE	0					
Philippines	tl_PH	0					
Poland	pl_PL	1	1	1	1	1	2443
Portugal	pt_PT	1	1	1	1	1	708
Romania	ro_RO	1	1	1	1	1	3077
Russian Federation	ru_RU	0					
Rwanda	fr_RW	0					
Senegal	fr_SN	0					
Serbia	sr_SP	0					
Slovak Republic	sk_SK	1	1	1	1	1	2147
Slovenia	sl_SL	1	1	1	1	1	2094
South Africa	en_ZA	1	1	1	1	1	2252
South Korea	ko_KR	0					
South Sudan	ar_EG	0					
Spain	es_ES	1	1	1	1	1	2502
Suriname	du_SR	0					
Sweden	sv_SE	1	1	1	1	1	8617
Switzerland	de_CH	1	0				
Tanzania	en_TZ	1	1	0			
Thailand	th_TH	0					
Togo	fr_TG	0					
Turkey	tr_TR	1	1	1	1	1	2214
Uganda	en_UG	0					
Ukraine	ru_UA	0					
UK	en_GB	0					
USA	en_US/es_US	0					
Uzbekistan	ru_UZ	0					
Venezuela	es_VE	0					
Viet Nam	vi_VN	0					
Zambia	en_ZM	0					
Zimbabwe	en_ZW	0					
Sub-TOTAL		34	29	25	23	19	66429
Wage-Indicator	english	1	1	1	0	0	4060
TOTAL							70489

Table 7-2 Overview of sources for the countries included in the Merged database

COUNTRY	SOURCE
Albania	http://www.akafp.gov.al/lista-kombetare-e-profesioneve/
Austria	Übersicht über die Gliederung der ÖISCO
Belgium	http://statbel.fgov.be/nl/statistieken/opendata/datasets/tools/nomenclaturen/
Bulgaria	http://www.nsi.bg/bg/content/261/basic-page
Czech Republic	https://www.czso.cz/csu/czso/klasifikace_zamestnani_-cz_isco-
Denmark	https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=16711&sid=disco
Estonia	http://metaweb.stat.ee/?siteLanguage=ee
Finland	downoad, Finnish/Swedish/English
Israel	retreived from researchers
Latvia	www.lm.gov.lv/upload/darba_devejiem/Profesiju_salidzinajums_gadi.xlsx
Lithuania	LIETUVOS RESPUBLIKOS PROFESIJŲ KLASIFIKATORIAUS ATNAUJINTA ŠEŠIAŽENKLĖ STRUKTŪRA PAGAL ISCO-08
Montenegro	downoad
Netherlands	https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs%20en%20beroepen/beroepenclassificatie--isco-en-sbc--/international-standard-classification-of-occupations--isco--/codelijsten-isco-08
Poland	Alfabetyczny indeks zawodów do KZiS (Dz. U. 28.08.14,poz.1145)st.22.12. 2014 systematic.pdf
Portugal	https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=107961853&PUBLICACOESmodo=2
Romania	http://www.fpen.ro/legislatie/COR.pdf
Slovak Republic	http://www.noveaspi.sk/products/lawText/1/75843/1/2
Slovenia	https://www.stat.si/doc/klasif/PretvornikSKPV2_SKP08_najnizjaRaven.xls
South Africa	http://www.statssa.gov.za/classifications/codelists/SASCO_2012.pdf
Spain	http://www.ine.es/jaxi/menu.do?L=1&type=pcaxis&path=%2Ft40%2Fcno11%2F&file=inebase
Sweden	http://www.scb.se/sv_/Dokumentation/Klassifikationer-och-standarder/Standard-for-svensk-yrkesklassificering-SSYK/
Turkey	http://tuikapp.tuik.gov.tr/DIESS/DosyaListeleAction.do?turId=1&tanimlayan_id=210&adi=ISCO-08