Deliverable Number: D8.7

Deliverable Title: Database of occupations for five languages + explanatory note

Work Package: WP8

Deliverable type: Other

Dissemination status: Public

Authors:

Kea Tijdens, University of Amsterdam/AIAS

Date Submitted: January 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

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

Executive summary

Many questionnaires have a question “What is your occupation” or similar. The answer is commonly asked as an open text field, challenging the survey holder to code the response into an occupation classification. Alternatively, in web surveys, respondents can self-identify their occupation from a database. Task 8.2 in SERISS aims to compile a database of occupations for 99 countries and 47 languages. Task 8.2 consists of five deliverables, all related to the database of occupations. Deliverable D8.7 ‘Database of occupations for five languages + explanatory note’ holds the database of 4,140 occupational titles, which have been translated in the five most spoken languages outside the EU28 area, notably Russian, Mandarin, Arabic, Hindi and Bahasa. Deliverable D8.7 also holds the translations of the ISCO-08 occupational structure for its 1 to 4 digits in these five languages. Both databases are available in the accompanying database **SERISS-Deliverable 8-7 translations five languages 2017**.

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

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. Moreover, 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 in order to improve survey measurement quality and 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. 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 WP 8 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), SHARE (CentERdata). Task 8.2 consists of five deliverables:

- D8.3 Database of occupations + explanatory note (M48)
- D8.4 Validation of ISCO-08 codes + explanatory note (M24)
- D8.5 Vacancy crawler and additions to database + explanatory note (M48)
- D8.6 Job task collector and additions to database + explanatory note (M48)
- D8.7 Database of occupations for five languages + explanatory note (M24)

This report concerns deliverable D8.7. The work for this deliverable was outsourced to the Institute for Employment Research (IER), University of Warwick, UK. The task related to the translation of the titles in the occupation database into Russian, Mandarin, Arabic, Hindi and Bahasa, and to a validation check of the coding. Two IER persons prepared the translations and conducted the validation, notably Rosie Day, Computing Officer, and Peter Elias, UK Strategic advisor for Data Resources for Social and Economic Research 2013-2018.

As agreed between SERISS and IER, the latter has undertaken the following activities:

1. Five classification files
   >> see accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017**), sheet ISCO-08 5dgt 5languages, consisting of the translated

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

occupational titles in the five languages, the English master label and the their ISCO-08 classifications;
>> see accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017)**, sheet ISCO-08 4dgt 5languages, presenting the translations of the ISCO-08 structure in the five languages

2. Performance test reports
   >> results can be found in Section 2
3. Disagreement reports
   >> see accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017**), sheet ISCO-08 code validation
4. Revised version of CASCOT software
   >> see IER - CASCOT webpage
   https://warwick.ac.uk/fac/soc/ier/software/cascot/internat/
5. Presentation of CASCOT software in workshop
   >> see accompanying power point presentation (**P.Elias_CASCOT.pdf**)

2. Explanatory note - database of occupations in five languages

Aims of Task 8.2

Task 8.2 aims to compile a database of occupational titles for 99 countries with in total 47 languages. All occupational titles are coded according to the International Standard Classifications of Occupations-08 (ISCO-08). The database of occupational titles initially is derived from the WageIndicator occupation database with 1,700 occupational titles used for 80 countries. Since 2000 the Netherlands WageIndicator web survey on work and wages used a database of occupational titles for respondent's self-identification (Tijdens 2015). Gradually this database was developed in a multilingual database, when more countries joined the survey from 2004 onwards. 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 ISCO88 to ISCO-08. By 2015, at the start of SERISS, the database held over 1,800 occupational titles from 80 countries, including translations in Russian, Mandarin, Arabic, Hindi and Bahasa. Approximately 100 of these 1,800 titles were used in one country only, and were not translated for other languages. Additionally, across countries the number of untranslated occupations varied largely, from over 500 to 0. The SERISS project allowed to expand and improve the database and to provide it to the research community on https://www.surveycodings.org/home .

As part of Deliverable D8.3 in Task 8.2 the number of the multilingual occupational titles in the initial database has been expanded to 4,000 titles, and the number of countries to 99. As part of Deliverable D8.4 in Task 8.2, validated, single-country occupational titles from ISCO-08 coding indexes from National Statistical Offices from more than 20 countries have been added to the database.

For Deliverable D8.7 in Task 8.2, partner IER has checked the database with the 4,000 occupational titles and collected, translated and validated the titles into the five largest languages outside the EU, notably Russian, Mandarin, Arabic, Hindi and Bahasa. The results of the disagreement check and the translations in the five languages are presented in the next sections.

Disagreement check

As part of the design of the occupational database, the lead partner checked the coding agreements across several sources: the codes used in ISCO-08 (ILO 2012), in the

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

WageIndicator database, and in the codes of Statistics Netherlands, abbreviated CBS (CBS 2013). For 181 job titles, this check resulted in different codes for the same job title. IER was asked to provide a final solution for the coding. All titles were then assigned a final code. The results of this can be found in the accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017)**.

Translations in the five languages

IER recruited translators for the translations. Partner University of Amsterdam had prepared translation instructions (see Appendix 1). The translators were all native speakers recruited through Warwick university recruitment agency and all had a minimum of an undergraduate degree. IER checked their translations of job titles against those suggested by Google Translate to ensure that they had not used this approach as their main source. Typically about 30% agreed with Google Translate.

The IER translators checked the Russian, Mandarin, Arabic, Hindi and Bahasa translations of the initial database of 1,700 titles, though not all titles had been translated. The results are presented in table 1, showing that the initial database had around 1,300 translated job titles. The results of the IER check varied largely. Whereas only 17 titles were proposed for a change by the Russian translator, this was 1144 for the Indonesian translator, as the table below shows. Note that it is not uncommon that the translation of occupational titles varies across translators (see SERISS report D8.4, Tijdens and Kaandorp, 2018).

Table 1 Number of initial translated job titles and the comparison with the IER translated job titles

Initial translation vs IER translation	hi_IN	ru_RU	zh_CN	ar_EG	ba_ID
Identical translations	562	1300	386	947	167
not identical translations	750	17	912	369	1144
Total	1312	1317	1298	1316	1311

The IER translators translated also the remaining 2,300 job titles, not yet translated. For Arabic and Bahasa Indonesian IER has also combined the SERISS index of job titles with a national index but the coding indexes were not used to expand the database. In total the database included 4,140 occupational titles. IER has asked the translators to check any duplicate job titles that appeared in more than one unit group to see which group was the best fit. Table 2 shows that in Hindi, Russian, and Arabic some 3% of the translated job titles were duplicate translations, indicating that two different job titles in the Master file had been translated similarly, because the language at stake does not differentiate between the two job titles. For Chinese and Indonesian the duplicate rate was around 1%. Across the five countries, the job titles that are classified as duplicates vary largely.

Table 2 Number of total translated job titles and their duplicates

Duplicate translations	hi_IN	ru_RU	zh_CN	ar_EG	ba_ID
Total job titles	4143	4143	4143	4143	4143
Of which duplicates	137	126	42	134	50
% duplicates	3.3%	3.0%	1.0%	3.2%	1.2%

The accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017**) presents the translations of the 4,140 job titles, whereby duplicates are left empty.

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

Translation check for the ISCO-08 structure in the five languages

Finally, the IER translators checked the 1 - 4 digit occupational groups in the ISCO-08 structure for the five languages. The accompanying database (**SERISS-Deliverable 8-7 translations five languages 2017)** presents the translations.

3. 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, Kaandorp C (2018) Validation of ISCO-08 codes + explanatory note. Deliverable D8.4 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

4. Appendix 1 Instructions for translators

The translation applies to a list of occupational titles, used in a web survey for the question: 'What is your occupation?'. A web visitor has to navigate the list to self-identify his/her occupation by means of semantic matching. This is similar to – for example – Google Search.

The list of occupations is drafted in English and needs to be translated in the specified languages.

- Please translate all occupational titles in singular (as is the case in English)
- If the translator has to choose between a literal translation and a translation that is used in the national labour market, please use the title reflecting the labour market reality.
- If one English title can be translated with two equivalent words, which both are used frequently, the two words can be included. See example below:

SOURCE LABEL	NLD LABEL
Shipbroker	Scheepsagent; Cargadoor

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

- However, in case of two equivalent words, where one word is used often and the other is used rarely, only the often used equivalent word should be included.
- In case in your language words are written together (e.g. eisenbahnhinandherschreiber in German) please use blanks in these words where possible, to facilitate reader's quick understanding
- In case two distinct English occupational titles are translated similarly in your language, this is not a problem. We will later remove similar titles.
- The word 'Manager' is used for people who manage a company of a department of substantial size.
- The word 'Engineer' is solely used for highly skilled occupations.
- The word 'Technician' is solely used for skilled occupations.
- The word '(certified)' is only added for highly skilled occupations.
- The word '(not certified)' is only added for skilled occupations.
- The word 'farmer' is used for people who are heading a farm with few staff. In case of a large, almost industrial farm, the occupation is called 'farm manager'.
- The words 'mechanic', 'machine operator', 'servicer', 'installer' are used solely for semi-skilled occupations.
- The word 'hand' – for example 'farm hand' is solely used for unskilled occupations.

When you don't know the occupation, you may the Occupation Quick Search of ONet, see https://www.onetonline.org/find/

For any questions, please email Kea Tijdens, University of Amsterdam, k.g.tijdens@uva.nl

*****

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