SERISS is funded by the EU-Horizon 2020 programme (nr 654221, July 2015-July 2019)

MEASURING OCCUPATIONS: RESPONDENT'S SELF-IDENTIFICATION FROM A LARGE DATABASE

Paper prepared for the European Conference on Quality in Official Statistics

Madrid, Spain, 31 May - 3 June 2016

Dr Kea Tijdens, University of Amsterdam / AIAS, Netherlands

Abstract

Most surveys use an open-ended question to measure occupational titles, followed by office coding. This is expensive and time-consuming, and some texts can be coded only at a highly aggregated level or not at all. Alternatively, in web-surveys or during a CAPI interview respondents can self-identify their occupation from a large database of coded occupational titles. For coding quality the size of the database is important, given that a national labour market easily comprises 10,000s job titles. This paper details the database.

For many years, the worldwide WageIndicator websites on work and wages has applied this self-identification method. In its Salary Check web-visitors can identify their occupation and view the related salaries. In its Salary Survey respondents are asked to self-identify their occupation. Both applications use the same multilingual database of approximately 1,600 occupational titles, all coded ISCO08 at five-digit level. Users can navigate the database through a 3-level search tree or through semantic matching. Nine out of ten use the latter.

As part of WorkPackage 8 of SERISS, the database is extended to 5,000 occupational titles, coded ISCO08 5-digit for 47 languages, using the coding indexes of the National Statistical Offices. These occupations are translated into English and their codes are compared. The subsequent solutions for 'same occupation–different code' problems are detailed.

An API (Application Programming Interface) is designed for the survey holders (free of charge during the SERISS project). For some countries, the database includes a gender filter showing (fe)male titles to (fe)male respondents. A life demo of the database will be shown.

WP8 facilitates an occupation>>industry prediction algorithm for respondents' easy self-identification of industry, as the majority of occupations are industry-bound. The API shows respondents a list of the five most likely industries, including an option 'other', which then allows respondents to search an industry database.

1 Introduction: measuring occupations

An inventory of 33 questionnaires from Europe and the United States, including those from international surveys as ESS and ECWS and national surveys as LFSs and SOEP, revealed that 25 surveys used an open text question and that only 14 asked for a job description (Tijdens, 2014a). Only in a few surveys the interviewer was asked to code the occupation during the interview, using a show card with the 2-digit occupational units, but all other surveys relied on office-coding. Office-coding is expensive and time consuming, and some entries can be coded only at highly aggregated levels or not at all. Survey agencies therefore mostly apply dictionary approaches for (semi-)automatic occupational coding. Recently, machine learning algorithms for a so-called auto-coder appear to be a promising development, but requiring a substantial amount of manually coded occupations to be used as training data. Unfortunately, this approach does not serve multi-country surveys, because auto-coders are language specific. In multi-country surveys each country typically conducts its own occupational coding and the coding instructions are the only source to ensure that the same job titles are coded similarly across countries. This paper explores self-identification as a means to collect occupational information in multi-country surveys. Table 1 defines the main concepts used.

Table 1 The main concepts used in this section

Self-identification	Survey respondents self-identify their occupational title from a look-up table
Look-up table	The list shown to the survey respondent, depending of the locale of the survey
Search tree	Tool facilitating respondents to browse the look-up table by means of a tree
Semantic matching	Tool facilitating respondents to browse the look-up table by means of typing the title in part or as a whole
Locale	The combination of language and country, for example en_US and es_US
Source list	The list of all occupational titles in the database, in English
Translation	The translation of an occupational title in the source list to the language of the respective locale (or vice versa the translation of the locale's title into English for the source list)
Entry	One occupational title in any of the look-up tables
Code	The ISCO-08 codes of the occupational titles in the source list
Database	The joint source list, codes and translations

2 Self-identification of occupation

CAWI and CAPI surveys allow respondents to self-identify their occupation during the interview, using a look-up table of occupational titles and a tool allowing respondents to browse the table. Of course, the occupational titles in the look-up table have to be coded according to an occupational classification. In this paper we fully adhere to ILO's International Standard Classification of Occupations (ISCO08).

For four reasons self-identification is advantageous over an open format question with office-coding. First, if designed well, the look-up table will consist only of occupations at the same level of aggregation. Second, unidentifiable occupational titles are absent. Third, field- or office-coding is not needed. Finally, in case of multi-country data-collections, survey operations and, in case of a multilingual database, the look-up will be comparable across countries.

However, for four reasons this method is disadvantageous. First, for respondents it is cognitively demanding to search their job title (Tijdens, 2014b). This is particularly the case when using a search tree only, but less so when a semantic matching tool is used. With Google and other search engines very widespread, semantic matching has become a familiar activity for many respondents. Second, by definition the look-up table is incomplete and therefore some respondents may not find their job title or are unable to aggregate it into an occupational title provided in the table. Third, it may be time-consuming for respondents to search for their job title. Finally, in mixed-mode surveys bias effects will occur when combining open format questions with closed format ones.

Given that any national labour market easily reaches 10,000s of job titles, the table should include a large number of occupational titles. Self-identification of occupation therefore poses high demands to the look-up table. This paper details first the experiences with self-identification of occupation in the WageIndicator Salary Survey and Salary Check and its Occupation Database. Then the paper details the progress of the database currently developed for 99 countries with 47 languages to be used for self-identification in national and international surveys.

3 Experiences with self-identification of occupation

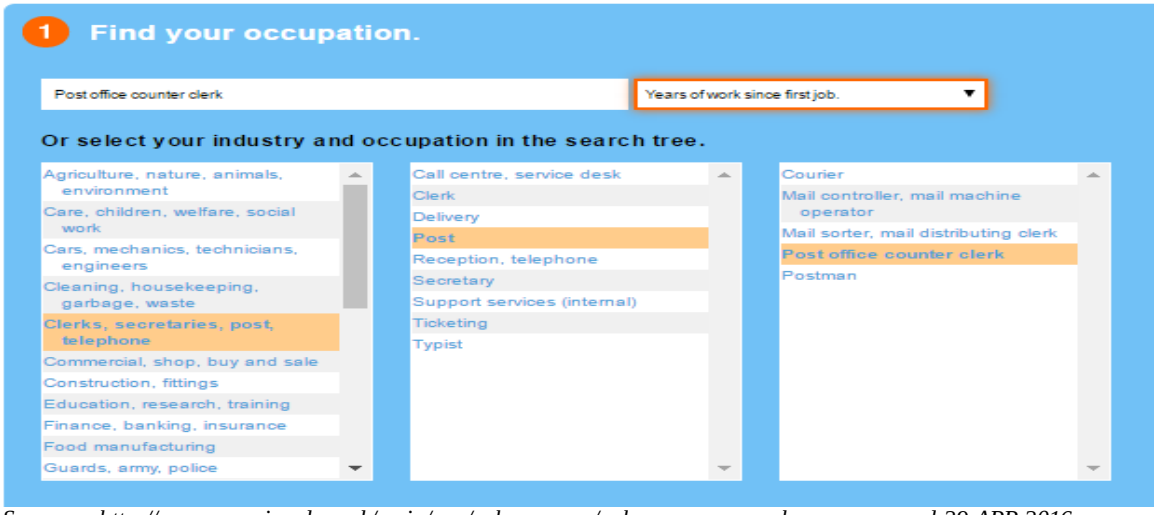
The first WageIndicator website on work and wages started in the Netherlands in 2001, and today WageIndicator is operational in 85 countries, receiving millions of web visitors (32 million in 2015). Each national website contains a Salary Check where web-visitors can

identify their occupation and view the related salary. Each national website has also a Salary Survey where respondents among others are asked to self-identify their occupation. The web survey consists of one standardized questionnaire, which is translated in the national language(s). The survey is posted continuously on all national websites, inviting web visitors to complete the survey. Hence, it is a volunteer survey. Both applications use the same multilingual database of currently 1,896 occupational titles, all coded ISCO08 at five digits. Users can navigate the database by means of a 3-level search tree or by a semantic matching tool.

In 2001 the occupation database started with a brief look-up table derived from the National Occupational Classification of Statistics Netherlands. In 2004, when thanks to the WOLIWEB project (2004-2006, EU-FP6 nr 506590), the web-survey expanded to seven other EU countries, some countries preferred to translate the Dutch list, others had occupation lists from their national statistical offices with their national coding. These lists were however not translated in English. On behalf of the self-identification of occupations the Salary Survey in each country had its own list and its own coding. The EurOccupations project (2006-2009, EU-FP6 nr 028987) aimed at drafting one occupation database for eight EU countries by merging the existing national look-up tables into one database of similar occupations in the countries at stake. By early 2007 the database held 1,433 occupational titles, derived from the ISCO88 coding index, yet restricted to occupations that were likely to have large numbers of jobholders. All titles were translated by professional translators and carefully checked by national labour market experts. In spring 2008 ILO published the final version of its ISCO-2008 classification and in June 2009 the draft definitions of the 4 digit occupation units (ILO, 2006). All titles in the database could then be coded ISCO08. Finally, the titles used in the EurOccupations similarity test of 150 occupations were included, resulting in a database of 1,594 occupational titles (Tables - EurOccupations Database of Occupations for 8 countries, 2009; Tijdens, 2010). National labour market experts associated with WageIndicator then translated the look-up table for countries other than the eight EU countries. By the end of 2009 the database was implemented in the WageIndicator Salary Survey, at the time covering 37 countries. Countries could add or remove occupations in their national look-up tables. For the UK, a number of management occupations were added, whereas for Germany skill levels were further detailed by distinguishing occupations at university and higher vocational level, and for the Czech Republic and Slovakia medical specialists were added. The source list of the current database (version 28, dated Sept 2015) includes 1,896 entries, of which 1,359 are translated for all 85 countries (42 languages) and the remaining occupational titles are translated for a few countries only.

Until the early 2010s, respondents of the Salary Survey could only browse their national look-up table by means of a 3-level search tree, as the screenshot in Figure 1 shows. As an alternative for a search tree, respondents in the web survey can now use a semantic matching tool, as Figure 2 shows. An occupation_API facilitates the Salary Surveys and the Salary Checks.

Figure 1 Screenshot of the search tree in the WageIndicator Salary Check



Source: http://www.paywizard.co.uk/main/pay/salarysurvey/salary-survey-employees, accessed 29 APR 2016

Figure 2 Screenshot of the semantic matching tool in the WageIndicator Salary Check

Source: http://www.paywizard.co.uk/main/pay/salarysurvey/salary-survey-employees, accessed 29 APR 2016

In 2015 almost 201,000 Salary Survey respondents and more than 296,000 Salary Check users from more than 80 countries used the occupation database. Unfortunately, we cannot identify what percentage used the search tree and what percentage used the semantic matching (time stamps to identify so will be implemented Summer 2016). Even with almost half a million web visitors using the tool last year, WageIndicator receives only once per two weeks a message 'my occupation is not in your list', indicating that the tool is functioning rather well.

4 Extending the Global Occupation Database

The SERISS project (2015-2019, EU-H2020 nr 654221) allows to extent the current database, as part of SERISS's Work Package 8 '*A coding module for socio-economic survey questions*'. The work package seeks to harness technological improvements and provide tools that can greatly increase the efficiency and reliability of the coding of open text responses in social

surveys. Occupation, industry, employment status, educational attainment, field of education, and social inclusion are core variables in most socio-economic and health surveys. However, their measurement, especially in a cross-cultural, cross-national and longitudinal context, is cumbersome, not sufficiently standardized and often quite expensive. The measurement of occupations through multi-country surveys faces two problems. First, a number of large countries have not adapted to the ISCO-08 classification (UK, USA, Germany, France, Italy, Poland) and maintain their own classification. For these countries no high quality ISCO08 coding index is available. Second, national survey agencies provide the office coding and thus international comparability is weak. The SERISS work package takes recent technological developments as an opportunity to improve survey measurement quality and provide cost-effective solutions to Research Infrastructures. As part of WP8, the global Occupation Database is enlarged to approximately 5,000 occupational titles, to 99 countries and to 47 languages. Survey holders can use the database for their national or multi-country surveys in CAWI or CAPI modes. This section details the principles underlying the drafting of the database.

A two-sided approach is used for the extension of the database. First, the source list of English occupational titles is expanded and these titles are translated into the 47 languages. Second, occupational titles from coding indexes from non-English countries are added after being translated into English. Both approaches will be explained hereafter.

On behalf of the first approach a source list of English occupational titles has been developed by merging four databases. In 2012 ILO published its final version of the ISCO-08 classification with 2,107 occupational titles (ILO, 2012). In 2013 ILO published a coding index in an excel file with three times as many occupational titles (7,150 titles). Next we merged the WageIndicator database with 1,896 titles, all coded ISCO08. Finally we merged a database of 284 occupational titles, taken from the UK SOC-2010 classification, which had been translated in 12 languages, as part of a WageIndicator project, and were recoded in ISCO-08. The merged file totalled to 11,306 entries, including duplicates. After removing duplicate entries and male/female duplicate titles a total of 8,635 entries remained. In a next step the entries with a high degree of similarity were identified and from these only the most common occupational title was included in the final source list. The final source list included 4,143 unique occupational titles. Finally, as the main purpose of the database is to serve self-identification of respondents in surveys, all entries were checked for phrasing that is typically applied in office coding but had to be rephrased for self-identification. For example 'Corporal, army' was changed into 'Army corporal' and 'Maker, accordion' was so into 'Accordion maker'. Only the latter serve self-identification, the former do not. If the WageIndicator database has no translations available, the entries will be translated by a professional translation agency into the 47 languages.

In the second approach coding indexes from national statistical offices were merged. The selected indexes were from non-English speaking countries, were coded ISCO08 and included more entries than the 433 4-digit occupational units. We collected 18 indexes from Albania, Austria, Bulgaria, Czech Republic, Denmark, Estonia, Finland, Latvia, Serbia, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and Turkey. These coding indexes were prepared for merging, whereby for example male/female titles in one entry were separated into two entries with a male and a female title. Then, using Google translate and a few dictionaries, all titles were translated into English. In a next step the

indexes were merged into one large database, whereby occupational titles with the same ISCO-08 code and the same English translation were considered the same occupation in the different languages. This database had more than 50,000 entries.

In a final step, the databases generated will be merged and used to populate the API. During the conference a live demonstration will be given.

5 An API for survey holders

An API (Application Programming Interface) is currently designed by SERISS partner SHARE-CeNTERDATA. Survey holders can register for the tool free of charge during the SERISS project. The related website will provide information about the use of the API, and about the countries and languages for which the API is available. Survey holders can include the tool in their web survey and they will be provided with the ISCO08 codes of the occupations of respondents. For CAPI interviews, a tool will be provided that interviewers can use to ask respondents to self-identify their occupation or use the tool themselves in order to instantly identify respondent's occupation.

6 Further plans

The deliverables list of SERISS WP8 includes three other topics of interest for survey holders.

First, to ensure quality measurement of education, this task aims to compile a country-specific database of educational attainment for the 99 countries, thereby extending GESIS' CAMCES database of educational qualifications to 99 countries, based on documentation from UNESCO and Eurostat and including a coding scheme into ISCED-2011. The database will be in in the national language(s) with data labels provided in English. To ensure quality measurement of field of education a general (not-country-specific) database of field of education and training will be developed for the languages at stake, thereby using existing – multilingual - databases such as from UNESCO, DISCO, ESCO, and similar. As is the case with the occupation_API, survey holders will be able to use the education_API in their web surveys and CAPI surveys.

Second, to ensure quality measurement of the industry variable, a multilingual database of industries will be developed by extending WageIndicator's industry database for 80 countries to 99 countries, including coding schemes into Nace-2.0 and ISIC-rev4. To facilitate the semantic matching, synonyms will be provided, among which a list of company names, as some respondents tend to provide the name of their company to the question 'In which industry do you work?'. Similar to the occupation_API, survey holders will be able to use the industry_API in their web surveys and CAPI surveys.

To facilitate advanced measurement of industries, an occupation – industry prediction is currently developed. This prediction allows survey holders to ask for industry with a restricted list of most likely industries, given the selected occupation, to provide an answer to the question 'In which industry do you work?'. Of course, the bottom of the list has an option 'other' allowing respondents to select another industry than the five listed, using a search tree. In the WageIndicator web survey non-response to the industry question is relatively high, most likely because industry is also asked by means of a search tree and a semantic matching tool, which is another cognitive demanding exercise for survey respondents. For this occupation >> industry prediction a merged, harmonized dataset of

more than 20 international surveys has been prepared. Currently merged dataset is analysed for these predictions.

Third, to facilitate quality coding of social stratification measures across countries, a database of employment status questions and answers (Q&A) will be prepared by extending WageIndicator's employment status Q&A for 80 countries to 99 countries, including a coding scheme to ISCE-93. Together with the occupation question, these questions allow the survey holder to measure respondents' socio-economic status according to international standardized classification systems. Similar to the occupation_API, survey holders will be able to use these survey questions in their web surveys and CAPI surveys.

References

ILO (2006) Draft classification structure – ISCO-08. Including Annex 1 for updating ISCO- 88 into ISCO-08. Geneva, International Labour Office

ILO (2012) International Standard Classification of Occupations ISCO-08 Volume 1 Structure, Group Definitions And Correspondence Tables, Geneva, International Labour Office (http://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_172572.pdf, accessed 17-12-2015).

Tijdens, K.G. (2010) Measuring occupations in web-surveys the WISCO database of occupations, University of Amsterdam, AIAS Working Paper 10-86, http://www.uva-aias.net/uploaded_files/publications/WP86-Tijdens.pdf

Tijdens, K.G. (2014a) Reviewing the measurement and comparison of occupations across Europe, University of Amsterdam, AIAS Working Paper 149.

Tijdens, K.G. (2014b) Drop-out rates during completion of an occupation search tree in web-surveys, Journal of Official Statistics, 30 (1), pp. 23–43.

****