Social Sciences & Humanities Open Cloud

Deliverable 3.3 Populating EQB, meeting DDI standards

Dissemination Level	PU
Due Date of Deliverable	28/02/2022 (M38)
Actual Submission Date	10/03/2022
Work Package	WP3 - Lifting Technologies and Services into the SSH Cloud
Task	T3.1 Multilingual Terminology
Type	Report
Approval Status	Approved by EC - 27 April 2022
Version	V1.0
Number of Pages	p.1 – p.32

Abstract:

This report describes the process of populating the European Question Bank with the multilingual online continuous WageIndicator Salary Survey, which is implemented in 160 countries around the world and can be completed in 58 languages. This report describes how all required question-level metadata was imported into the EQB. Three challenges were encountered: importing question grids, referring to APIs, and handling language-country locales.

History

Version	Date	Reason	Revised by
0.1.	24/01/2022	Draft for peer review	Janna Besamusca & Kea Tijdens
0.2	11/02/2022	Peer Review	Tomas Kabina
0.3	18/02/2022	Revised Draft	Janna Besamusca & Kea Tijdens
0.4	20/02/2022	WP leader review	Daan Broeder
1.0	25/02/2022	Review and minor editorial edits	Janna Besamusca & Kea Tijdens

Author List

Organisation	Name	Contact Information
CLARIN ERIC	Janna Besamusca	J.W.Besamusca@uu.nl
CLARIN ERIC (WageIndicator Foundation (WWI))	Kea Tijdens	K.G.Tijdens@uva.nl

Executive Summary

The European Question Bank (EQB) aims to provide a central search facility across all the Consortium of European Social Science Data Archives' (CESSDA) survey holdings. It uses a question-level metadata schema based on the DDI-Lifecycle standard (DDI Alliance 2020). The ambition is for users to be able to find survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title. The population of the EQB with the multilingual online continuous WageIndicator Salary Survey (Tijdens & Osse) provides an important use case to test the feasibility of harvesting social science questionnaires for the EQB. The survey, started in 2000 in the Netherlands, is currently collecting data in 160 countries around the world and can be completed in 58 languages. For each country, the questions have been translated.

This report describes how all required question-level metadata, including survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title, was imported into the EQB. Due to the existence of structured metadata, the extraction of all relevant questionnaire texts did not pose any challenges. Conversion of the structured metadata to DDI lifecycle standards required by the EQB was performed using conversion scripts that can be re-used in the future. The metadata published in the beta version of the online EQB effectively gives users access to survey questions, their translations to applicable languages, answer categories, pre- and post-question texts.

Three main challenges were encountered in the process of populating the EQB with the multilingual WageIndicator Salary Survey. Firstly, the introduction of question grids into the EQB posed challenges to the search function since the search terms most likely to be used by end users were not found in the question texts but answer categories. This was solved by expanding the search function and by breaking down the relevant question grids into separate questions. Secondly, the inclusion of answer categories including extremely long lists of potential answers, which are commonly stored outside the survey XML sheet and called on using APIs, posed challenges to the inclusion of all answer categories in the EQB as well as to the display of those answer options to end users. In the EQB, access to answer categories stored in APIs and their translations is now provided by the inclusion of links to the source API. This solution ensures users have access to the most up-to-date versions of the ontologies and vocabularies, as well as to all translations. Thirdly, the large number of language versions of the WageIndicator version made high demands on the IT system, which has to be able to correctly process numerous writing systems, as well as to the user-friendliness of the web interface. The beta version of the EQB is capable of processing all included language systems. The translations for the 217 language versions are collapsed by default in the EQB to make the site more user friendly; users can click to expand and access all translations.

Abbreviations and Acronyms

API	Application Programming Interface. A computing interface defining interaction between separate programs. APIs often allow client software to interact with server applications.
DDI	Data Documentation Initiative
DOI	Digital Object Identifier System
EOSC	European Open Science Cloud
EQB	European Question Bank
CESSDA	Consortium for European Social Science Data Archives
FAIR	FAIR (Findable, Accessible, Interoperable, Re-usable)
XML	Extensible Markup Language. An open standard markup language defining a set of rules for documents that are both human- and machine-readable.

Table of Contents

1. Introduction and background

In the SSHOC Project Task 3.1 (Multilingual terminology) several efforts were undertaken to develop and make accessible a number of resources and tools designed to meet the needs of a pan-European SSH infrastructure. Making data and services accessible and usable in SSH is very much a matter of providing relevant translations, translation of metadata, multilingual vocabularies, terminology extraction across languages, multilingual databases. Partners involved in this task developed and supported a number of such resources and tools in four topical areas: (1) the Multilingual Question Bank (EQB), (2) Multilingual document collections that enable content searches across different languages, (3) Multilingual metadata and taxonomies that permit metadata-based discovery using different languages, and (4) the creation of female job titles for the multilingual occupation ontology.

This paper reports on the first topic, the Multilingual Question Bank (EQB), which was led by CLARIN ERIC (WWI) and CESSDA ERIC. The European Question Bank, which is being prepared for release at the time of writing of this report, aims to enrich the collected surveys with respect to translation quality, provenance, and the use of infrequent words. The EQB uses a question-level metadata schema based on the DDI-Lifecycle standard (DDI Alliance 2020). For the development of the EQB the services and software of Colectica have been engaged (Colectica 2021a, 2021b). The Colectica Platform offers solutions for data owners and data archives that are looking to increase the expressiveness and longevity of the data collected through standards-based metadata documentation. The aim is to provide a central search facility across all the Consortium of European Social Science Data Archives' (CESSDA) survey holdings. The ambition is for users to be able to find survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title.

For this deliverable, survey questions from the multilingual WageIndicator Salary Survey were prepared for populating the multilingual European Question Bank (EQB), maintained by CESSDA, using the identifiers needed to meet DDI Lifecycle standards. In so doing, FAIR data principles were a central concern. Adding the WageIndicator Survey to the EQB makes it easier for users to find and gain free access to the survey questions, answer categories and pre- and post-question texts in all 217 language versions of the survey. Furthermore, due to the required preparation of DDI standard meta data for the inclusion of the WageIndicator Salary Survey in the EQB, the interoperability is increased. The inclusion of the WageIndicator Salary Survey adds value to the EQB by making translations of common social science survey questions into 58 languages accessible to for re-use under a creative commons (CC01) license.

Finally, the WageIndicator Salary Survey serves as a use case to explore and test the feasibility of harvesting multilingual questionnaires into an open access social science question bank. This report is centered around the experiences of that use case and focused on describing the process of populating the EQB with the multilingual WageIndicator Salary Survey, and on the challenges encountered in this process. After introducing the survey instrument and the documentation system used by the

WageIndicator Foundation, section 5 describes the steps followed to populate the EQB: the extraction of survey texts, conversion to DDI standards, and the visualization in the EQB web interface, of which a beta version is currently in existence. In section 6, three challenges are discussed that related to the inclusion of question grids, multilingual ontologies and vocabularies, and the 217 language versions of the questionnaire. Section 7 summarizes the lessons learned.

2. The WageIndicator Survey on Wages and Working Conditions

WageIndicator Foundation (WIF) is an NGO in the Netherlands, that operates frequently visited national websites with work-related content and web tools about frequently searched topics, such as collective labour agreements. The Foundation started in 2001 with one website in the Netherlands. By 2020 it had 40 million visitors worldwide and websites in more than 150 countries, among which all EU27 countries. WageIndicator has a team of more than hundred persons, located in many countries in all continents, and maintains the WageIndicator continuous online survey on wages and working condition, or Salary Survey.

A multilingual online survey

The WageIndicator Salary Survey is a continuous online, multilingual, multi-country self-completion survey. The survey started in 2000 in the Netherlands. Since 2004, websites have been launched in many European countries, in North and South America and in countries in Asia. From 2008 on web sites have been launched in more African countries, as well as in Indonesia and in a number of post-Soviet countries. Currently, the survey is implemented in 160 countries around the world and can be completed in 58 languages. For each country, the questions have been translated. Multilingual countries employ multilingual questionnaires.

The web-survey generates cross-sectional data and might be used for longitudinal analyses using repeated cross-sections. The survey has detailed questions about earnings, benefits, working conditions, employment contracts and training, as well as questions about education, occupation, industry and household characteristics. The WageIndicator Survey includes regularly extra survey questions for project targeting specific countries, for specific groups or about specific events.

The WageIndicator data is derived from a volunteer survey, inviting the millions of annual web-visitors to the national WageIndicator websites to complete the web-survey. As such, the WageIndicator Survey is a non-probability sample, or so-called convenience sample. Comparisons with representative studies found an underrepresentation of male labour force, part-timers, older age groups, and low educated persons.

Data documentation and administration

The WageIndicator Survey on Wages and Working Conditions is archived at the Research Data Center of the IZA Institute of Labor Economics (Tijdens & Osse). The data is organised in annual releases. The data of the period 2000-2005 is released as one dataset. Each data release consists of a dataset with

continuous variables and one with project variables. The continuous variables can be merged across years. All variable and value labels are in English. The data does not include the text variables and verbatims form open-ended survey questions, these are available in Excel-Format upon request. Access to the data is provided free of charge to non-for-profit research, replication and teaching purposes, after submitting an access request.

Question-level metadata was maintained internally but did not conform to DDI lifecycle standards. However, WageIndicator did document structured metadata using an xml sheet, which is integrated with the web environment of the survey. The xml file containing the WageIndicator question-level metadata contains structured meta data across a number of columns, as is displayed in the data extract in figure 1.

	SURVEY	UNIQUE IDENTIFIER (PER SURVEY)	ITEM TYPE	INITIAL TRANSLATIONS	TEXT FOR TRANSLATORS	T
8	SALARY SURVEY	sourceapi_countryr	API source			
9	SALARY SURVEY	regihome_API	question - API chooser	You live in the region of	In which region do you live?	Y
0	SALARY SURVEY	sourceapi_regihome	API source			
1	SALARY SURVEY	nrofjob1	question	Do you have a paid job	Do you have a paid job?	Y
2	SALARY SURVEY	nrofjob1_hint	hint	As an employee, self-employed or othe	As an employee, self-employed or o	Y
3	SALARY SURVEY	nrofjob1_1	option	Yes	Yes	N
4	SALARY SURVEY	nrofjob1_0	option	No	No	N
5	SALARY SURVEY	countryw_API	question - API chooser	You work in	In which country do you work?	Y
6	SALARY SURVEY	sourceapi_countryw	API source			
7	SALARY SURVEY	countryw_API_hint	hint	If you work at sea, please tick this in the	If you work at sea, please tick this i	Y
8	SALARY SURVEY	cobself_API	question - API chooser	You were born in	In which country were you born?	Y
9	SALARY SURVEY	sourceapi_cobself	API source			
0	SALARY SURVEY	countrye_API	question - API chooser	You were educated in	In which country did you receive yo	Y
1	SALARY SURVEY	sourceapi_countrye	API source			
2	SALARY SURVEY	yybirth	question - YY drop down	When were you born?	When were you born?	Y
3	SALARY SURVEY	sex	question	What is your gender?	What is your gender?	Y
4	SALARY SURVEY	sex_1	option	Woman	Woman	N
5	SALARY SURVEY	sex_2	option	Man	Man	N
6	SALARY SURVEY	PAK_sex	question	What is your gender?	What is your gender?	N
7	SALARY SURVEY	PAK_sex_1	option	Woman	Woman	N
8	SALARY SURVEY	PAK_sex_2	option	Man	Man	N
9	SALARY SURVEY	PAK_sex_3	option	Transgender	Transgender	N
0	SALARY SURVEY	PAGE_employed	heading	Your job	Your job	Y
1	SALARY SURVEY	contst71	question	Are you mainly...	What are you mainly?	Y
2	SALARY SURVEY	contst71_hint	hint	If you have more than one job, please a	If you have more than one job, plea	Y
3	SALARY SURVEY	contst71_140	option	Employee	Employee	Y
4	SALARY SURVEY	contst71_141	option	Civil servant	Civil servant	Y
5	SALARY SURVEY	contst71_202	option	Self-employed without employees, free	Self-employed without employees,	Y
6	SALARY SURVEY	contst71_201	option	Self-employed with employees	Self-employed with employees	Y

FIGURE 1 EXCERPT OF THE STRUCTURED DATA DOCUMENTATION FILE WAGEINDICATOR

Survey

The first column of the xml sheet contains the title of the survey instrument. For the purpose of this deliverable, this is always the Salary Survey.

Unique item identifier

The unique item identifier is a persistent identifier of each row. Unique identifiers linked to one question have the same root, which is the identifier of the question. Answer options, pre- and post-question texts use the identifier of the question as the root, followed by "_hint" or the code of the answer option. For example, the question "what is your gender" has the root identifier "sex". For the answer option "woman", which takes the value of 1 in the dataset, the unique identifier is "sex_1". These unique item identifiers are matched with IDs in the survey xml structure, such that correct translations are shown at right place.

Item type

Item type signals the nature of the item described in the row. Included item types are:

- Alert: alerts are messages that are triggered when invalid answers are given in the questionnaire. They can be messages like "Please enter a number higher than 0." or "Answering this question is required.".
- Heading: signals the start of a new topic in the questionnaire. Headings include, for example "your job", "your place of work" or "your personal background"
- Hint: hints are prompts included in the questionnaire, usually as pre-question text. These include texts like "Please select the main reason." Or "If you have worked both full-time and part-time alternately, please answer with reference to the most recent period."
- Question: rows with the item type 'question' contain the question text for the questionnaire
- Question – API chooser: in rows labeled 'Question – API chooser' the question text is not followed by rows with answer options. Instead, API questions draw on vocabularies outside the xml sheet for answer options. For example, the "regihome_API" question draws on a list of thousands of geographical regions and selects only those regions linked to the respondent's country of residence for display in the online questionnaire.
- Question – text box: this row contains a question for which no answer options exist. Respondents type their text-based answer in a text box in the online questionnaire.
- Question – number box: this row contains a question for which no answer options exist. Respondents type their numerical answer in a text box in the online questionnaire.
- Question – YY drop down: this row contains a question for which the answer options are not included in the xml sheet. Instead, the survey includes a standard drown down menu displaying a list of years,
- Question – MM drop down: this row contains a question for which the answer options are not included in the xml sheet. Instead, the survey includes a standard drown down menu displaying a list of 12 months,
- Option: rows containing the answer options of a question. Each row contains a single answer option. Thus, for a yes/no question, two rows of options are included: one for the "yes" answer and one for the "no" answer.

- Questionnaire text: text displayed at the start and end of the survey. This text is not linked to any questions and thanks respondents for completing the survey.

The item type column also contains a number of item types associated with question grids (batteries of questions), which are referred to as matrix groups. Matrix questions contain a battery of question items displayed on the same page in the online questionnaire and usually sharing one overarching question and one set of answer options. For example, the matrix group "firmpast" asks respondents "in the past 12 months, what has happened in your organization?" and lists a range of question items (firmdis1, firmpri2, etc.) that respondents answer "yes" (firmpast_yn_1) or "no" (firmpast_yn_0) to.

- Matrix group: the label matrix group signals that the following questions are part of a matrix question.
- Matrix question: rows containing the questions of a matrix group
- Matrix option: rows containing the answer options for a matrix group.

TABLE 1 EXAMPLE OF QUESTION GRID IN THE WAGEINDICATOR XML FILE

firmpast	matrix group	In the past 12 months, what has happened in your organization?
Firmdis1	matrix question	Staff redundancies
firmpri2	matrix question	Privatisation
firmreo1	matrix question	Reorganisation, organisational change
firmprot	matrix question	Employee protest, strike
firminit	matrix question	New strategic initiatives
firmreti	matrix question	… announced voluntary retirement?
Firmmerg	matrix question	Merger with another organisation
firmcom1	matrix question	Renewal of computer equipment / machinery
firmstra	matrix question	Competent management strategy
firmbro1	matrix question	Urgent threat of bankruptcy
firmpast_yn_1	matrix option	Yes
firmpast_yn_0	matrix option	No
firmpast_-8	matrix option	Not applicable

INITIAL TRANSLATIONS

The initial translations column contains the text of question, hints, alerts and answer options as displayed in the web survey.

VALUES OF VAR

The values of var (variables) column contains the values of answer options used in the dataset, where applicable.

VAR/VALUE LABEL

The variable)value label column contains the text labels of variables and values (answer options) as used in the dataset, where applicable.

VARNAME_dataset

The varname dataset column contains the name of variables as used in the dataset, where applicable.

TRANSLATIONS

Finally, a set of columns is used for translations to the 58 languages in which they survey is implemented. One column is created per survey 'locale', which consists of a two-letter indication of the language used, followed by a two-letter code for the country of the survey. For example, nl_BE and fr_BE represent the columns containing the Dutch language and French language versions of the Belgian questionnaire.

FIGURE 2 EXCERPT OF TRANSLATION COLUMNS FROM THE WAGEINDICATOR SURVEY METADATA XML

3. Populating the EQB with the multilingual WageIndicator Salary Survey

To populate the EQB with the multilingual WageIndicator Salary Survey, a number of steps were taken: survey questions, answer options, pre- and post-question texts were extracted, these texts were converted to question level meta-data in accordance with DDI lifecycle standards used by the EQB, and the resulting metadata were displayed in the EQB web interface, of which a beta version is currently being tested. In this section, each step is described in detail.

Extracting questionnaire texts

First, the survey questions and answer options needed to be extracted from the questionnaire texts and stored in a structured format. This step did not present any major challenges, due to the structured metadata format in which WageIndicator documents its survey, as was presented in the previous section. Thanks to the use of the xml file by WageIndicator, access to all 217 versions of the questionnaire (160 countries, 58 languages) could be provided in one file. To prepare the WageIndicator questionnaire texts for conversion to question-level meta data that can be harvested by the EQB, all columns that contained no relevant metadata and pertained solely to the display of text on the WageIndicator websites were removed. All remaining columns were checked for consistency of labels and for completeness.

Conversions to DDI standards

Secondly, the WageIndicator xml file with structured metadata needed to be harvested for the EQB, for which the metadata need to be converted to DDI standards. For this step, the services of the Colectica team were engaged, who prepared a harvesting script in cooperation with the WageIndicator and CESSDA-ERIC teams (Colectica 2021a). The script maps the contents of the original WageIndicator xml file to DDI lifecycle and saves a DDI Lifecycle 3.3 XML file that can be loaded into Colectica Designer and Colectica Repository for inclusion in the EQB. The following paragraphs outline how the data were converted to DDI standards and which adaptations were made to the WageIndicator structured metadata to allow for harvesting into the EQB. Table 1 summarizes the steps that were taken for each original metadata tag; additions and transformations to the original XML file done are described in the next section on challenges that were encountered during the conversion.

In the mapping, all questionnaire texts, including questions, answer options, pre- and post-question texts, are extracted from the column **initial translations**, which contains the texts from the English language master version of the survey. The column **survey**, which contains the name of the survey instrument, is used as the instrument label. This links each question to they study title, which the EQB aims to include. The column **unique identifier** contains the item name. All items linked to a question

have the same root, thus enable the creation of question-level metadata. The tag in the column **item type** indicates how the row is processed in the mapping. Tags in this column distinguish between questions, answer categories and pre- and post-question texts.

- Rows with the item type **heading** are used to create sequences within the survey instrument. The use of sequences allows users to browse the questionnaire by topic.
- Rows with the item type **question** are used to create a question item. Item type questions – api chooser, question – text box, question – number box, question – YY drop down, and question – MM drop down are treated in the same way. These questions are added to the survey instrument and the appropriate sequence.
- Rows with the item type **option** create a code and category and add it as a response option to the current question item. Thus, the mapping links each row containing a response option to an appropriate question.
- Rows with the item type **hint** are recognized as containing pre- or post-question texts. The text is added as an instruction to the current question item.
- Rows with the item type **matrix group** create a question grid, where the text in the row with the signifier 'matrix group' is treated as the question.
- Rows with the item type **matrix question** create a dimension and add it to the current question grid.
- Rows with the item type **matrix option** create a response option and link it to the current question grid.
- Rows with the item type **questionnaire text** create a statement and add it to the instrument and sequence. These texts, which include statements thanking respondents from completing the survey, are not linked to specific questions.
- When the item type is **alert**, the row is ignored, as these rows merely contain pop-up warnings for invalid values.

TABLE 2 TREATMENT OF WAGEINDICATOR TAGS FOR HARVESTING INTO THE EQB

WageIndicator meta data	EQB script	Added/Transformed for file harvesting
Survey	Survey instrument label	
Unique item identifier	Item name	
Item type	Indicates how the row is processed	
Alert	Ignore	
Heading	Used for grouping questions	
Hint	Included as pre-question text	
Question	Create question item and add it to the survey instrument and sequence	
Question – API chooser	Create question item and add it to the survey instrument and sequence	Addition of a row containing the source of the API the survey

		draws on the complement missing information for the user.
Question – text box	Create question item and add it to the survey instrument and sequence	
Question – number box	Create question item and add it to the survey instrument and sequence	
Question – YY drop down	Create question item and add it to the survey instrument and sequence	
Question – MM drop down	Create question item and add it to the survey instrument and sequence	
Option	Added to current question item as response option	
Questionnaire text	Create a statement, add to survey instrument and sequence	
Matrix group	Create question grid item	Manual addition of rows containing the signifiers "matrix group start" and "matrix group end" before and after the first and last rows linked to the matrix grid.
Matrix question	Create dimension, add to current question grid	
Matrix option	Create response option, add to current question grid	
Initial translations	Contains questionnaire texts (questions, answer options, pre- and post-question texts)	
Dataset type	Determines whether to create a represented variable	
Values of var	Represented variable values	
Var/value label	Represented variable label and value labels	
Varname dataset	Represented variable item name	
Translations	Text in all columns with language codes (locales) is processed as questionnaire text (translations) in the relevant language	Addition of buttons in the EQB user interface to collapse and expand translation options.

In the final part of the mapping, once the item type column has been used to determine the processing of each row and create question-level meta data, data in the remaining columns is used to add information about represented variables and translations to the identified questions. The column **DATASET TYPE** was used to determine whether the question was included in the dataset as a variable. If yes, a represented variable was created. The name of the variable was extracted from the column

VARNAME_dataset, variable and value labels from the column **VAR/VALUE LABEL**, and variable values of each answer option from the column **VALUES OF VAR**.

Finally, translations were linked to all questionnaire texts from the 217 columns containing the translated text for each locale (i.e., language and country combination). The translation columns include the translation of questions, answer option, pre- and post-question texts. All languages in the XML sheet were included in the conversion.

Display in the EQB environment

The final step in the population of the EQB with multilingual WageIndicator Salary on Wages and Working Condition was the visualization in the EQB web interface. The aim was to design the web interface in such a way that users can find survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title. In this remainder of this section, this web interface is discussed, with the aid of visuals from the current beta version of the web interface.

Figure 3 shows the landing page of the WageIndicator study in the European Question Bank (beta version). From this page, users can browse the full study by clicking on the button "192 questions", explore the questions per section sequences by clicking on the button "26 question groups", or see the full structure of the survey questions by clicking on the top button "Salary Survey Questions".
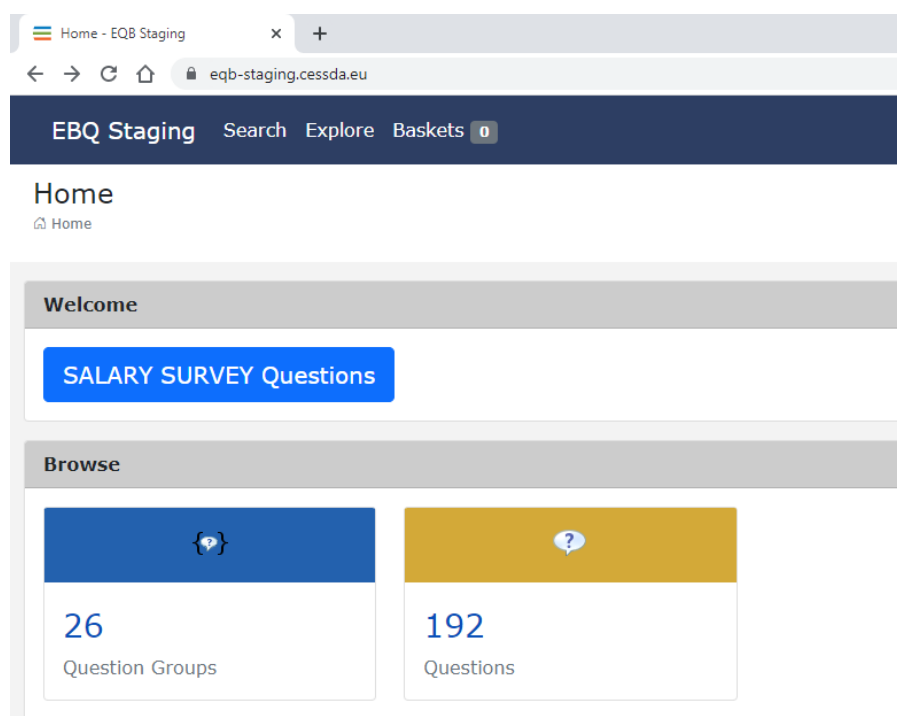
FIGURE 3 LANDING PAGE OF THE WAGEINDICATOR SURVEY IN THE EQB

When users use to browse the question groups, they are led to page where all sequences are listed, as displayed in Figure 4.
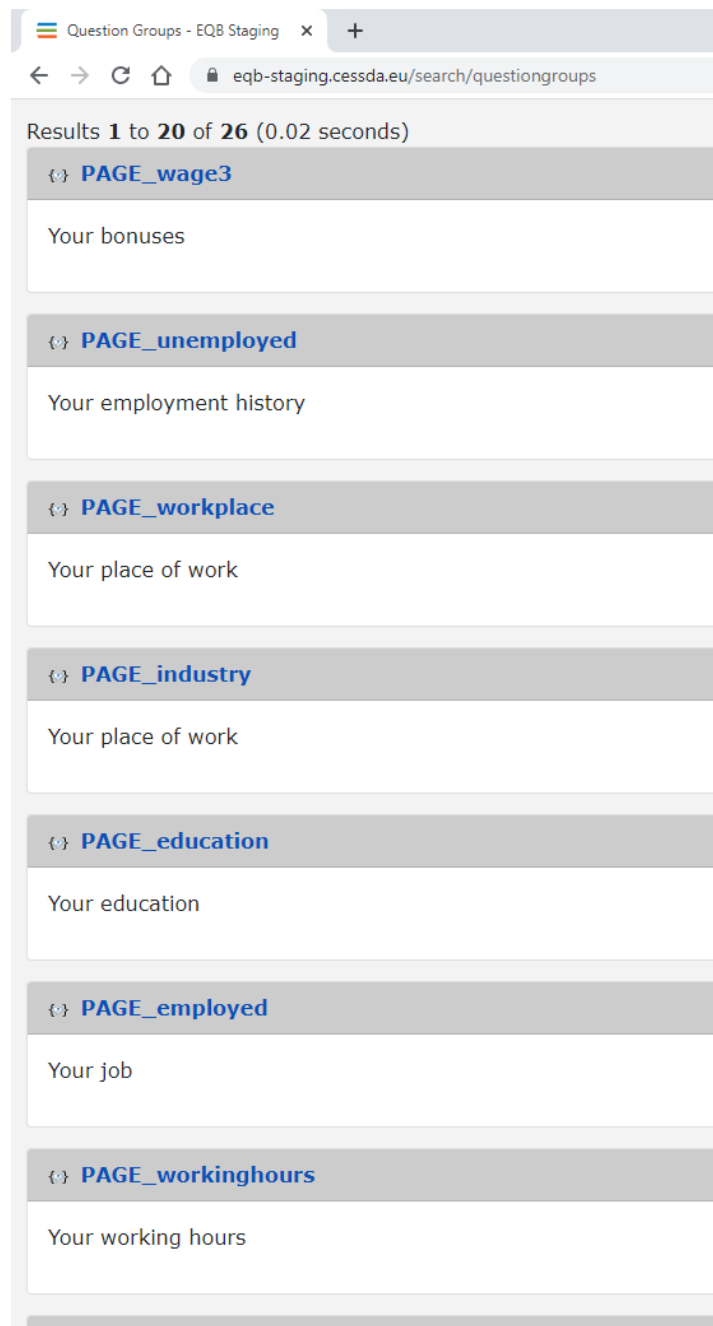
FIGURE 4 PAGE OF QUESTION GROUPS IN THE WAGEINDICATOR STUDY

When users choose to browse the questions instead of question groups, they are taken to the search tab (Figure 5). This page displays a search field as well as the list of all questions in the survey, including one-line previews for each question. As Figure 6 shows, question-level meta data is displayed when clicking

on any question. Displayed content includes the question identifier, which is equal to the variable name in the dataset if applicable, the question and answer categories. The name of the study and sequence that the question is included in can be seen under the heading "Question groups". The heading "concept" shows whether the question is included as a variable in the dataset and, if so, under which name. In this example, languages are not shown by default. The list of translations of the question text can be expanded by clicking on "show more languages". Translations of the response options can be found by clicking on the response option text. Figures 7 and 8 show how the translations are displayed in the EQB web interface.
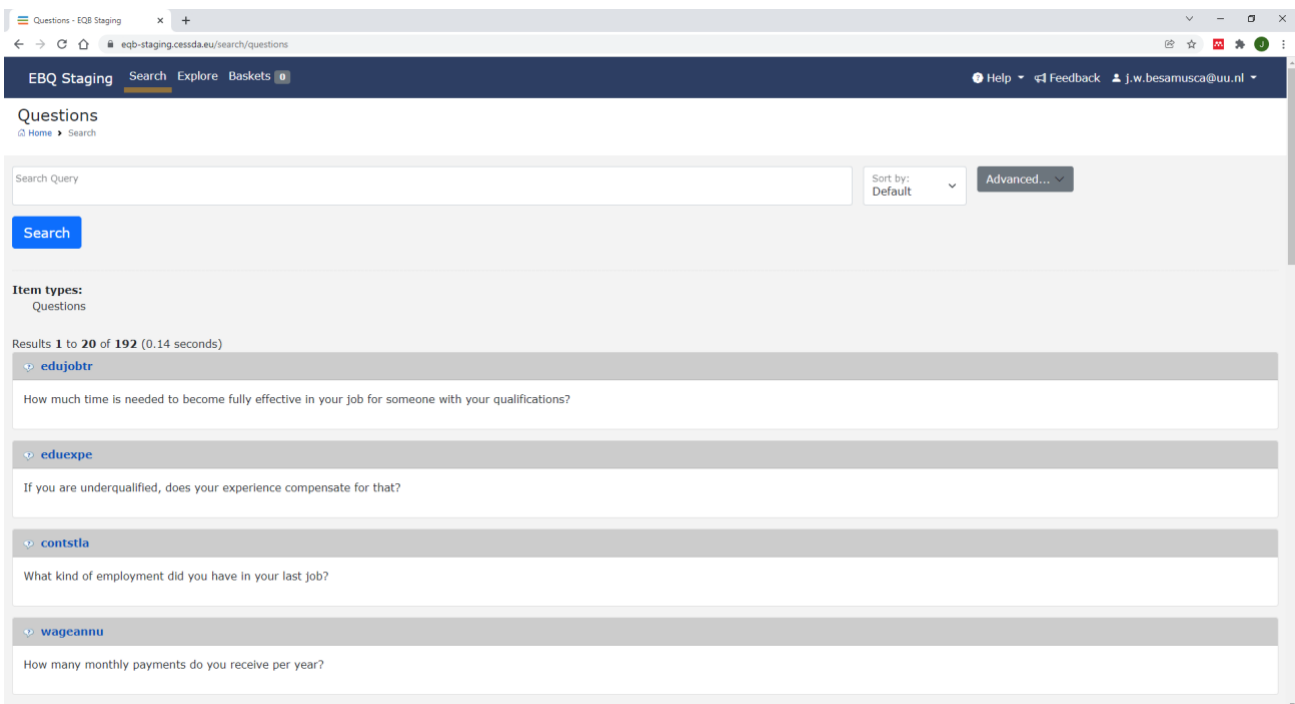
FIGURE 5 PAGE TO SEARCH WAGEINDICATOR SURVEY BY QUESTION
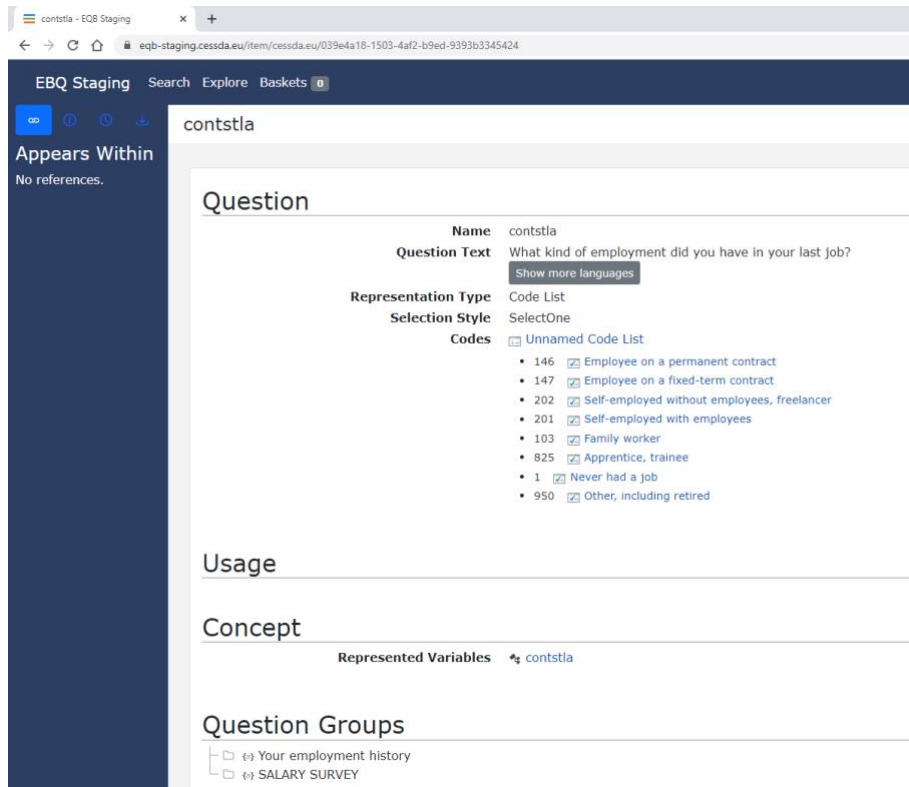
FIGURE 6 QUESTION-LEVEL META DATA DISPLAYED IN THE EQB WEB INTERFACE
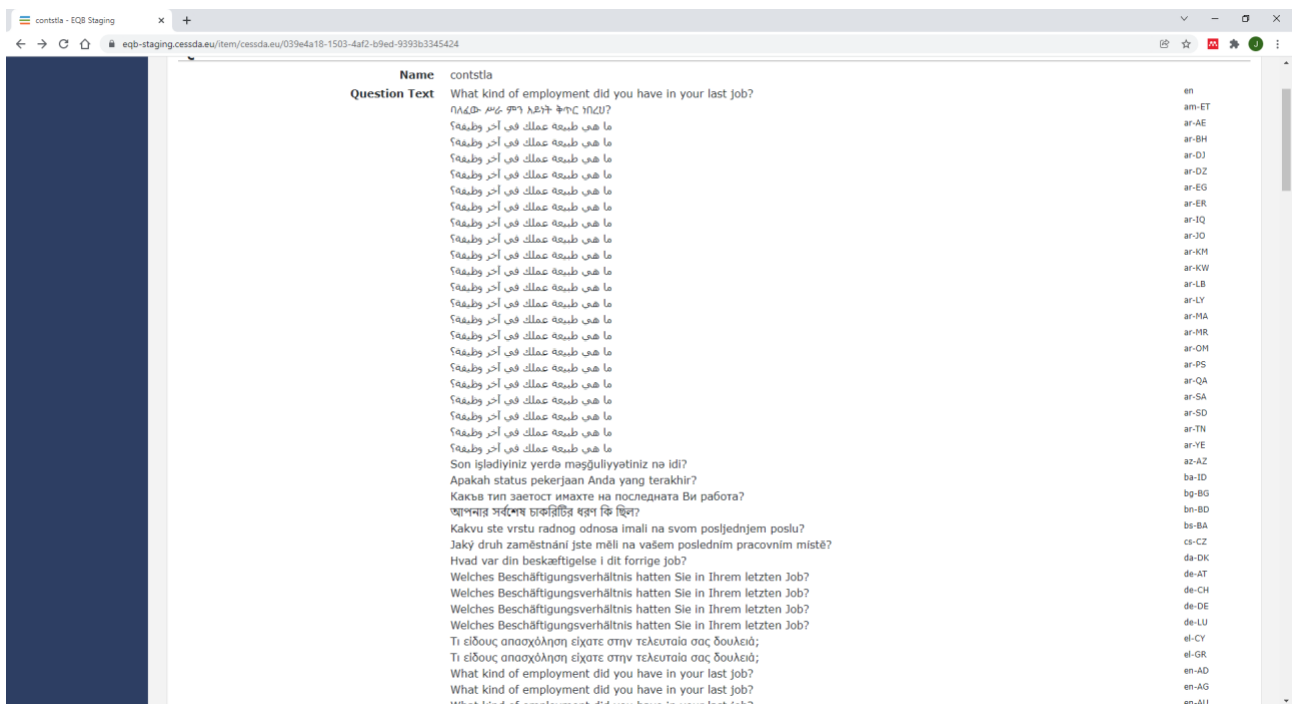
FIGURE 7 QUESTION TRANSLATIONS DISPLAYED IN THE EQB WEB INTERFACE
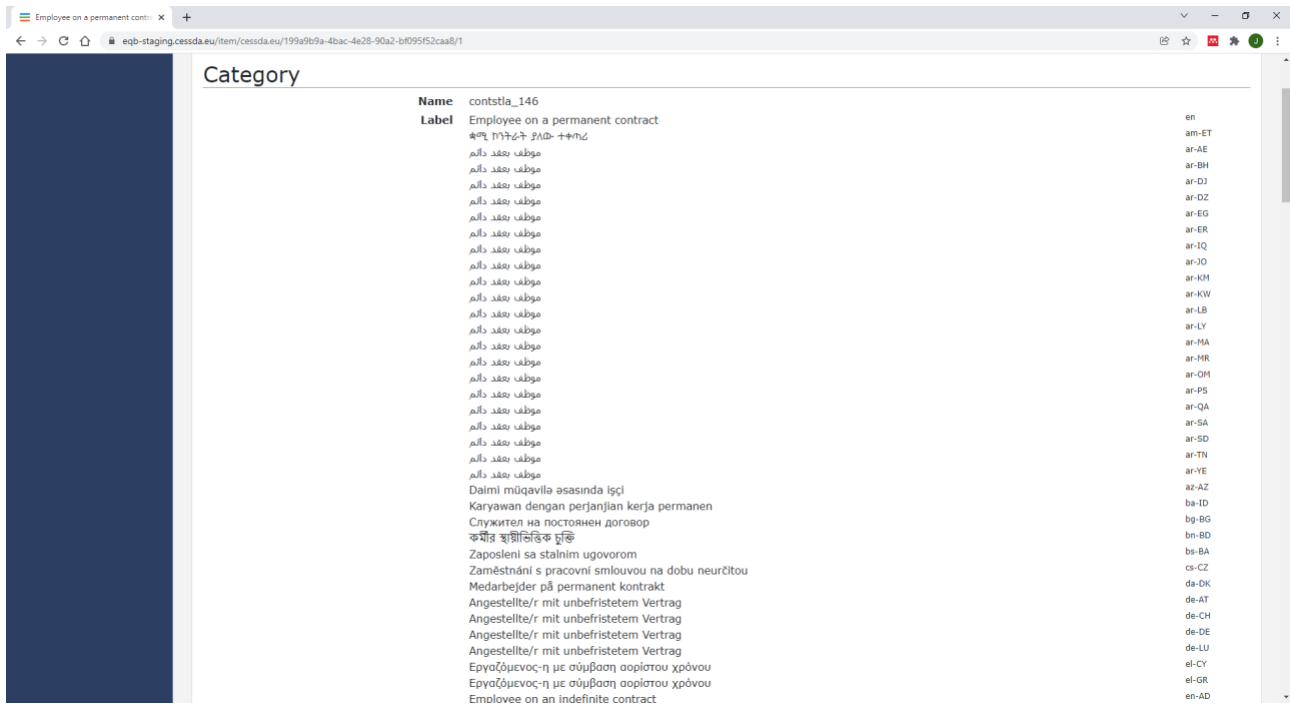
FIGURE 8 ANSWER CATEGORY TRANSLATIONS DISPLAYED IN THE EQB WEB INTERFACE

4. Challenges in the population of the EQB with the multilingual WageIndicator Salary Survey

In the process of populating the EQB with the multilingual WageIndicator Salary Survey, a number of challenges were encountered. Firstly, the introduction of question grids into the EQB posed challenges to the search function. Secondly, the inclusion of answer categories including extremely long lists of potential answers, which are commonly stored outside the survey XML sheet and called on using APIs, posed challenges to the inclusion of all answer categories in the EQB as well as to the display of those answer options to end users. Thirdly, the large number of language versions of the WageIndicator version made high demands on the IT system, which has to be able to correctly process numerous writing systems, as well as to the user-friendliness of the web interface. Each of these three challenges in described below.

Question groupings and search terms

The treatment of one particular group of item type, those related to question grid, presented a number of challenges, as also recognized in deliverable 9.5 of this project (Feasibility report on setting up a collection on questionnaires relating to ethnic and migrant minorities in the EQB) (Morales, Saji & Winn 2021). Question grids, or matrix questions, are popular tools in social science questionnaires that group a battery of questions on the same topic. The questions in the grid often aim to measure a single latent dimension (e.g., attitudes towards gender equality) and commonly share one instruction text or overarching question (e.g., "to what extent to you agree or disagree with the following statements") and one set of response options (e.g., strongly disagree, disagree, neither agree nor disagree, agree, strongly agree).

In the population of the EQB, question grids pose challenges because the overarching question often contains no more than the instruction text (e.g., to what extent do you agree or disagree with the following), whereas the substantive question is found in the individual items. However, when users search the EQB they will likely use search terms related to the substantive topic of interest rather. In reference to the FAIR data principles underlying the EQB, question grids risk being less 'findable' if the question grid is harvested as if it were one question.

Several solutions were explored, both in the preparation of the WageIndicator files for harvesting to the EQB and in the web environment of the EQB itself. The first solution is to expand the search function to cover not just questions but also answer categories. This solution is currently being explored with the ICT firm that is building the web environment, Colectica. The second solution is to adapt the question grids in the WageIndicator XML sheet before it is harvested into the EQB environment to treat the sub-items of the question grids as individual questions.

This second solution was also applied. All question grids where the terms likely to be used in user searches were found in the sub-items, and where those sub-items could be considered as stand-alone questions, were adapted. Individual questions were created for each sub-question, the answer categories of the question grid were multiplied for each new question, and the overarching instruction text of the grid was linked to the sub-item as a pre-question text (hint). As is shown in table 3, for example, the questions wrk_leave, wrk_sick and wrk_student are all followed by the same pre-question text (hint) and by the same answer categories (option). When making the new separate questions from the old question grids, integrating the instruction text and the sub-item question text into one fluent question was considered. However, this option was rejected due to the multilingual nature of the questionnaire. While changing the text in the English language questionnaire would be feasible, all 217 translations would be erroneous as a result of doing so.

TABLE 3 EXAMPLE OF REFORMATTED QUESTION GRID

wrk_leave	question	On child-care leave or other leave
wrk_leave_hint	hint	Are you currently...
wrk_leave_1	option	Yes
wrk_leave_0	option	No
wrk_leave_-8	option	Not applicable
wrk_sick	question	Unable to work due to long-term illness or disability
wrk_sick_hint	hint	Are you currently...
wrk_sick_1	option	Yes
wrk_sick_0	option	No
wrk_sick_-8	option	Not applicable
wrk_student	question	In full-time education / student
wrk_student_hint	hint	Are you currently...
wrk_student_1	option	Yes
wrk_student_0	option	No
wrk_student_-8	option	Not applicable

APIs- Including long lists of answer categories and data from outside the XML sheet

A second challenge was the inclusion of answer categories consisting of extremely long lists of potential answers. The social sciences commonly rely on ontologies and vocabularies to measure social constructs like occupations, educational classifications, sectors of industry, geographical regions, and religions. Predefined response categories are used in social science questionnaires like the WageIndicator Salary Survey, and can consists of hundreds, thousands or even tens of thousands of response categories.

Shared multilingual ontologies and vocabularies, such as those described in task 3.2 on multilingual ontologies and vocabularies of the SSHOC project (Martens & Tijdens 2021). Multilingual classified sets of items are presented a structured interface and stored in a relational database that surveys like the WageIndicator survey draw on using APIs. Using the shared ontologies, the web survey either allows respondents to find their answer category by typing the first few letters and selecting from the remaining answer categories, or to present only the relevant options. In the formal case, for example, a respondent asked what their occupation is may type 'teacher' into the answer field and then select from a short list of answer categories like teacher in primary education, secondary education, etc. In the latter case, a respondent educated in Norway, will choose their highest completed education degree from a set of Norwegian education tracks.

In the WageIndicator XML, questions drawing on externally kept vocabularies and ontologies are indicated as "question – API chooser". APIs are used to answer questions regarding the respondent's country (of residence, work, etc.), highest educational attainment, field of education, industrial sector, currency of the wage, occupation, and language of the interview/questionnaire (see table 4).

TABLE 4 WAGEINDICATOR QUESTIONS USING AN API

Language	question - API chooser	You want to answer the survey in
countryr_API	question - API chooser	You live in
regihome_API	question - API chooser	You live in the region of
countryw_API	question - API chooser	You work in
cobself_API	question - API chooser	You were born in
countrye_API	question - API chooser	You were educated in
educat_API	question - API chooser	What is the highest level of education you have attained?
educast_API	question - API chooser	At what stage of education are you at the moment?
educaf_API	question - API chooser	What is your main field of education?
occupai3_API	question - API chooser	What is your occupation?
u_occupai3_API	question - API chooser	What was your occupation?
firmow01_API	question - API chooser	Please select the country where the foreign investment comes from.
sector2_API	question - API chooser	What is the main activity of the organization where you work?
u_sector2_API	question - API chooser	What is the main activity of the organization where you worked?
wagecurr_API	question - API chooser	You received your last wage in [currency]
s_wagecurr_API	question - API chooser	You report your income in [currency]
occupafr_API	question - API chooser	What was your first occupation?

SVK_locfrjb	question - API chooser	Where was this job located?
regibir1_API	question - API chooser	Where were you born?
cobmothe_API	question - API chooser	In which country was your mother born?
cobfathe_API	question - API chooser	In which country was your father born?
language_API_f2f	question - API chooser	Language of interview

In the context of the EQB these ontologies and vocabularies present two challenges: (1) they are stored outside the WageIndicator XML and (2) including the answer categories in the EQB web interface would make the answer categories unreadable, this diminishing accessibility. The first challenge is easily overcome, as the answer categories stored in the APIs can be downloaded and harvested into the EQB. However, doing so would present two problems. First, as APIs are continuously updated EQB end users would get access to potentially outdated APIs. Second, the answer categories are stored in APIs exactly because they are too vast for a respondent to process during a questionnaire; they are therefore also too vast for most EQB end users.

To solve the accessibility issue of the missing answer categories without reducing the usability of the EQB, a decision was made to include the link to the source API in the question-level metadata in the EQB. Users search for a question for which answer categories are stored in an API, will thus be provided with the link to the API source. Via the link, they will be able to access the answer categories via a link that contains the up-to-date ontologies and vocabularies including the translations to all the languages of the WageIndicator Salary Survey.

Handling language locales and translations

The third challenge that was encountered in the process of populating the EQB with the WageIndicator Salary Survey relates to the large number of language versions. The WageIndicator survey can be completed in 58 languages, which are shown in table 5. While the majority of these languages use a Latin script and read from left to right, not all included languages do. The language versions of the questionnaire include writing systems that require questions to be displayed from right to left, as well as character systems and different scripts (for example Cyrillic, Amharic, Arabic, Hindi, etc.). To incorporate all language versions of the WageIndicator questionnaire into the EQB, the software therefore needs to be able to process a large variety of characters, signs and scripts. This requirement was discussed with the ICT firm developing the EQB web interface and in the beta version of the EQB, all languages and scripts are included.

TABLE 5 LANGUAGES INCLUDED IN THE WAGEINDICATOR XML SHEET

Albanian	Georgian	Romanian
Amharic	German	Russian
Arabic	Greek	Serbian

Armenian	Hebrew	Sinhala (Sri lankan)
Bahasa (Indonesian)	Hindi	Slovakia
Bengali	Hungarian	Slovenian
Bosnian	Icelandic	Somali
Bulgarian	Italian	Spanish
Burmese	Japanese	Swahili
Chinese	Khmer	Swedish
Croatian	Korean	Tagalog
Czech	Lao	Thai
Danish	Latvian	Turkish
Dutch	Lithuanian	Turkmen
Dzongkha	Malay	Ukrainian
English	Mongolian	Urdu
Estonian	Nepali	Uzbeki
Farsi (Persian)	Norwegian	Vietnamese
Finnish	Polish	
French	Portuguese	

A second challenge relates to the visualization of the different language versions of the questionnaire. While the questionnaire can be completed in 58 languages, there are many more language versions of the questionnaire because questionnaires exist in localized versions. WageIndicator questionnaires are linked to so-called 'locales' which constitute a combination of the country of residence of the respondent and the localized language version in which the questionnaire was completed: for example, in Belgium the Flemish questionnaire is identified as nl_BE and the Walloon questionnaire as fr_BE. Translations into a single language may differ slightly per country, such as between Indian, UK and US English or Portuguese and Brazilian. Therefore, the language versions of the questionnaire cannot be reduced to the number of languages but is equal to the number of locales. This means that 217 translations of the questionnaire are imported into the EQB.

While the initial intent was to display all translations in the EQB, the volume of language versions makes this impractical for users. Therefore, translations are now collapsed by default and buttons were added to expand the translations.

5. Conclusion

The European Question Bank (EQB) aims to provide a central search facility across all the Consortium of European Social Science Data Archives' (CESSDA) survey holdings. It uses a question-level metadata schema based on DDI-Lifecycle standard. The ambition is for users to be able to find survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title. The population of the EQB with the multilingual online continuous WageIndicator Salary Survey provides an important use case to test the feasibility of harvesting social science questionnaires for the EQB.

This report describes how all required question-level metadata, including survey questions, their translations to applicable languages, answer categories, pre- and post-question texts, and the study title, was imported into the EQB. Due to the existence of structured metadata, the extraction of all relevant questionnaire texts did not pose any challenges. Conversion of the structured metadata to DDI lifecycle standards required by the EQB was performed using conversion scripts that can be re-used in the future. The metadata published in the beta version of the online EQB effectively gives users access to survey questions, their translations to applicable languages, answer categories, pre- and post-question texts.

Three main challenges were encountered in the process of populating the EQB with the multilingual WageIndicator Salary Survey. Firstly, the introduction of question grids into the EQB posed challenges to the search function since the search terms most likely to be used by end users were not found in the question texts but answer categories. This was solved by expanding the search function and by breaking down the relevant question grids into separate questions. Secondly, the inclusion of answer categories including extremely long lists of potential answers, which are commonly stored outside the survey XML sheet and called on using APIs, posed challenges to the inclusion of all answer categories in the EQB as well as to the display of those answer options to end users. In the EQB, access to answer categories stored in APIs and their translations is now provided by the inclusion of links to the source API. This solution ensures users have access to the most up-to-date versions of the ontologies and vocabularies, as well as to all translations. Thirdly, the large number of language versions of the WageIndicator version made high demands on the IT system, which has to be able to correctly process numerous writing systems, as well as to the user-friendliness of the web interface. The beta version of the EQB is capable of processing all included language systems. The translations for the 217 language versions are collapsed by default in the EQB to make the site more user friendly; users can click to expand and access all translations.

6. References

Colectica (2021a). Colectica. Colectica 6.2 Documentation. Retrieved November 12, 2021, from https://docs.colectica.com/.

Colectica (2021b). Colectica Designer. Colectica - Colectica Designer. Retrieved November 12, 2021, from https://www.colectica.com/software/designer/.

DDI Alliance. (2020). Getting Started with DDI: Version 2 | Data Documentation Initiative. DDI Alliance. https://ddialliance.org/resources/getting-started/v2

Martens, M., & Tijdens, K. (2021). D3.4 Multilingual ontologies for Occupation, Industry, Regions and cities, Food items, and Religion, with use case (V1.1). Zenodo. https://doi.org/10.5281/zenodo.4724837

Morales, L., Saji, A., & Winn, M. (2021). D9.5 Feasibility report on setting up a collection on questionnaires relating to Ethnic and Migrant Minorities in the European Question Bank (1.0). Zenodo. https://doi.org/10.5281/zenodo.5898536

Saji, A., & Morales, L. (2020). D9.4 Database with the metadata of surveys to EMMs across Europe (v1.0). Zenodo. https://doi.org/10.5281/zenodo.4558307

Tijdens, K.G. and P. Osse, WageIndicator continuous web-survey on work and wages. Amsterdam: University of Amsterdam/AIAS and Stichting Loonwijzer. doi: https://dx.doi.org/10.15185/wif.1

List of Tables

List of Figures

Appendix - List of national websites of the WageIndicator Salary Survey

COUNTRY	CONTINENT	LANGUAGE	SALARY SURVEY
Albania	Europe	Albanian	Rrograime.al
Algeria	Africa	Arabic	Rawateb.org/Algeria
Angola	Africa	Portuguese	Meusalario.org/Angola
Argentina	South America	Spanish	Elsalario.com.ar
Australia	Oceania	English	Mywage.org/Australia
Austria	Europe	German	Lohnspiegel.org/Osterreich
Azerbaijan	Europe	Azerbaijani	Qazancim.az
Azerbaijan	Europe	Russian	Mojazarplata.az
Bahamas	North America	English	Mywage.org/Bahamas
Bahrain	Asia	Arabic	Rawateb.org/Bahrain
Bangladesh	Asia	Bengali	Mywage.org.bd
Belarus	Europe	Russian	Mojazarplata.by
Belgium	Europe	Dutch	Loonwijzer.be
Belgium	Europe	French	Votresalaire.be
Benin	Africa	French	Votresalaire.org/Benin
Bolivia	South America	Spanish	Tusalario.org/Bolivia
Bosnia and Herzegovina	Europe	Bosnian	Mojaplata.ba
Botswana	Africa	English	Mywage.org/Botswana
Brazil	South America	Portuguese	Meusalario.org.br
Bulgaria	Europe	Bulgarian	Moiatazaplata.org
Burkina Faso	Africa	French	Votresalaire.org/BurkinaFaso
Burundi	Africa	French	Votresalaire.bi
Cabo Verde	Africa	Portuguese	Meusalario.org/CaboVerde
Cambodia	Asia	English	Prake.org
Cambodia	Asia	Khmer	Prake.org
Cameroon	Africa	French	Votresalaire.org/Cameroun
Canada	North America	French	Fr.mywage.ca
Canada	North America	English	Mywage.ca
Central African Republic	Africa	French	Votresalaire.org/Centrafricaine
Chad	Africa	French	Votresalaire.org/Tchad

Chile	South America	Spanish	Tusalario.org/Chile
China	Asia	Chinese	Wageindicator.cn
Colombia	South America	Spanish	Tusalario.org/Colombia
Congo	Africa	French	Votresalaire.org/Congo
Congo Brazzaville	Africa	French	Votresalaire.org/CongoBrazzaville
Costa Rica	North America	Spanish	Tusalario.org/Costarica
Côte d'Ivoire	Africa	French	Votresalaire.org/Cotedivoire
Croatia	Europe	Croatian	Mojaplaca.org
Cuba	North America	Spanish	Tusalario.org/Cuba
Cyprus	Europe	Greek	Mywage.org/Cyprus
Czech Republic	Europe	Czech	Mujplat.cz
Denmark	Europe	Danish	Lontjek.dk
Dominican Republic	North America	Spanish	Tusalario.org/RepublicaDominicana
Ecuador	South America	Spanish	Tusalario.org/Ecuador
Egypt	Africa	Arabic	Rawateb.org/Egypt
El Salvador	North America	Spanish	Tusalario.org/Elsalvador
Eswatini	Africa	English	Mywage.org/Eswatini
Estonia	Europe	Estonian	Mysalary.ee
Ethiopia	Africa	Amharic	Mywage.org/Ethiopia-am
Ethiopia	Africa	English	Mywage.org/Ethiopia
Fiji	Oceania	English	Mywage.org/Fiji
Finland	Europe	Finnish	Wageindicator.fi
France	Europe	French	Votresalaire.fr
Gabon	Africa	French	Votresalaire.org/Gabon
Gambia	Africa	English	Mywage.org/Gambia
Germany	Europe	German	Wageindicator.de
Ghana	Africa	English	Mywage.org/Ghana
Ghana	Africa	English	Africapay.org/Ghana
Greece	Europe	Greek	Mywage.gr
Guatemala	North America	Spanish	Tusalario.org/Guatemala
Guinea	Africa	French	Votresalaire.org/Guinee
Guinea-Bissau	Africa	Portuguese	Meusalario.org/GuinéBissau
Haiti	North America	French	Votresalaire.org/Haiti
Honduras	North America	Spanish	Tusalario.org/Honduras
Hungary	Europe	Hungarian	Berbarometer.hu
India	Asia	English	Paycheck.in

Indonesia	Asia	Indonesian	Gajimu.com
Iraq	Asia	Arabic	Rawateb.org/Iraq
Ireland	Europe	English	Mywage.org/Ireland
Italy	Europe	Italian	Iltuosalario.it
Jamaica	North America	English	Mywage.org/Jamaica
Japan	Asia	Japanese	Kyuryocheck.jp
Jordan	Asia	Arabic	Rawateb.org/Jordan
Kazakhstan	Europe	Russian	Mojazarplata.kz
Kenya	Africa	English	Mywage.org/Kenya
Kenya	Africa	English	Africapay.org/Kenya
Kuwait	Asia	Arabic	Rawateb.org/Kuwait
Latvia	Europe	Latvian	Mysalary.lv
Lebanon	Asia	Arabic	Rawateb.org/Lebanon
Lesotho	Africa	English	Mywage.org/Lesotho
Liberia	Africa	English	Mywage.org/Liberia
Lithuania	Europe	Lithuanian	Mysalary.lt
Luxembourg	Europe	French	Lohnspiegel.lu
Luxembourg	Europe	German	Lohnspiegel.lu
Madagascar	Africa	French	Votresalaire.org/Madagascar
Malawi	Africa	English	Mywage.org/Malawi
Malaysia	Asia	Malay	Gajimu.my
Mali	Africa	French	Votresalaire.org/Mali
Malta	Europe	English	Mywage.org/Malta
Mauritania	Africa	Arabic	Rawateb.org/Mauritania
Mauritius	Africa	English	Mywage.org/Mauritius
Mexico	North America	Spanish	Misalario.org
Moldova	Europe	Romanian	Salariulmeu.org/Moldova
Mongolia	Asia	Mongolian	Mysalary.mn
Montenegro	Europe	Bosnian	Mojazarada.me
Morocco	Africa	Arabic	Rawateb.org/Morocco
Mozambique	Africa	Portuguese	Meusalario.org/Mocambique
Myanmar	Asia	Burmese	Mywage.org/Myanmar
Namibia	Africa	English	Mywage.com/Namibia
Netherlands	Europe	Dutch	Loonwijzer.nl
New Caledonia	Oceania	French	Votresalaire.org/NouvelleCalédonie
New Zealand	Oceania	English	Mywage.co.nz

Nicaragua	North America	Spanish	Tusalario.org/Nicaragua
Niger	Africa	French	Votresalaire.org/Niger
Nigeria	Africa	English	Mywage.ng
North Macedonia	Europe	Albanian	Mojataplata.mk
Oman	Asia	Arabic	Rawateb.org/Oman
Pakistan	Asia	English	Paycheck.pk
Palestine	Asia	Arabic	Rawateb.org/Palestine
Panama	North America	Spanish	Tusalario.org/Panama
Papau New Guinea	Oceania	English	Mywage.org/PNG
Paraguay	South America	Spanish	Tusalario.org/Paraguay
Peru	South America	Spanish	Tusalario.org/Peru
Philippines	Asia	English	Mywage.org/Philippines
Poland	Europe	Polish	Twojezarobki.com
Portugal	Europe	Portuguese	Meusalario.pt
Puerto Rico	North America	Spanish	Tusalario.org/PuertoRico
Qatar	Asia	Arabic	Rawateb.org/Qatar
Réunion	Africa	French	Votresalaire.org/LaReunion
Romania	Europe	Romanian	Salariulmeu.org/Romania
Russia	Europe	Russian	Mojazarplata.ru
Rwanda	Africa	English	Mywage.org/Rwanda
Saudi Arabia	Asia	Arabic	Rawateb.org/SaudiArabia
Senegal	Africa	French	Votresalaire.org/Senegal
Serbia	Europe	Serbian	Mojazarada.rs
Sierra Leone	Africa	English	Mywage.org/SierraLeone
Slovakia	Europe	Slovak	Mojplat.sk
Slovenia	Europe	Slovene	Mojaplaca.si
South Africa	Africa	English	Mywage.co.za
South Korea	Asia	Korean	Mywage.org/Korea
South Sudan	Africa	English	Mywage.org/SouthSudan
Spain	Europe	Spanish	Tusalario.es
Sri Lanka	Asia	English	Salary.lk
Sudan	Africa	Arabic	Rawateb.org/Sudan
Suriname	South America	English	Mywage.org/Suriname
Sweden	Europe	Swedish	Lonecheck.se
Switzerland	Europe	French	Votresalaire.org/Suisse
Switzerland	Europe	German	Lohnspiegel.org/Schweiz

Tanzania	Africa	English	Mywage.org/Tanzania
Tanzania	Africa	English	Africapay.org/Tanzania
Tanzania	Africa	Swahili	Mywage.org/Tanzania-sw
Togo	Africa	French	Votresalaire.org/Togo
Trinidad and Tobago	North America	English	Mywage.org/TrinidadandTobago
Tunisia	Africa	French	Votresalaire.org/Tunisie
Tunisia	Africa	Arabic	Rawateb.org/Tunisia
Turkey	Europe	Turkish	Maascetveli.org
Uganda	Africa	English	Mywage.ug
Uganda	Africa	English	Africapay.org/Uganda
Ukraine	Europe	Ukrainian	Mojazarplata.com.ua
Ukraine	Europe	Russian	Mojazarplata.com.ua
United Arab Emirates	Asia	Arabic	Rawateb.org/UAE
United Kingdom	Europe	English	Wageindicator.co.uk
United States of America	North America	English	Paywizard.org
United States of America	North America	Spanish	Tusalario.org/USA
Uruguay	South America	Spanish	Tusalario.org/Uruguay
Venezuela	South America	Spanish	Tusalario.org/Venezuela
Vietnam	Asia	Vietnamese	Luong.com.vn
Yemen	Asia	Arabic	Rawateb.org/Yemen
Zambia	Africa	English	Mywage.org/Zambia
Zanzibar	Africa	Swahili	Mywage.org/Zanzibar
Zimbabwe	Africa	English	Mywage.org/Zimbabwe