Creating a Question Bank for Music Survey Data Harmonisation

Daniel Antal, CFA Anna Márta Mester

2024-11-09

Table of contents

> **Note**
>
> This document is a very early stage documentation of our work. It is intended for first consultation and statement of work, and must be considered as work-in-progress.

Summary

A question bank aims to create a useful database of survey questions and answer options. Surveying aims to collect structured data that can be processed with statistical means, and therefore it seldom allows respondents to provide free-ranging answers. Responses collected to the questions then can be processed into statistical variables. For example, answers to

the question `What is your age?` provide a numeric representation of the data subject's age, which is then processed into various variables like the `mean age of the respondents` or `median age of the respondents`, or minimum and maximum ages[1].

The aim of survey harmonisation is to arrive to comparable 'mean age of respondents" variables from different surveys. This means that in two or more surveys, numeric answers representing the respondent's age are averaged with the same arithmetic algorithm. The numeric representation of the age may be available from administrative and questionnaire surveys, and a question bank focuses on different questions that can yield a harmonised representation of answers.

A question bank must identify synonyms and translations of the same question. For example, `What is your age` can be translated to Hungarian as `Mi az életkorod?` (informal), `Mi az Ön életkora` (formal) and `Hány éves vagy?` (literally: how many years do you have). From a semantic point of view, they are identical questions if they are answered by numbers in the same numeric range. However, they cannot be considered synonymous or translated to the question `What is your birth year?`, even if with further arithmetic we can infer the answer to original question.

Numbers are perhaps the least ambiguous concepts represented by the answers, although the age of the person as a number is a somewhat fuzzy concept (What about a person who just turned 32 and somebody who will turn 33 in a month?).

Most survey questions want to capture other concepts that need a harmonised conceptualisation in the form of a pre-agreed codebook or scale. For example, the answers to the `What is your gender?` question may be understood in a four-category structure („female", „male", „other", „decline") allowing for refusal to answer the question. The same question (i.e., synonyms or translations) will not result in comparable variables if one questionnaire uses a binary female-male coding and another uses a female-male-other or even more nuanced structure. Therefore we must add the representation of the potential answer range in all the languages in which we want to ask the questions. A translation of only the question is not adequate, because both the respondent and the interviewer (in case the survey is not self-administered) must understand both the question and the accepted structure of the answers.

Last, but not least, the same statistical procedure must be applied on harmonised response items. For example, category frequencies or averages must handle missing values identically among samples. If one variable omits declined answers, then it can only be harmonised with a variable representing another survey if it also omits declined answers.

Retrospecitve or ex-post harmonisation usually allows us to create new variables with the application of novel statistical formulas on responses given to the same question in different questionnaires. If we can compare the questions and the semantic coding of the responses,

[1]This research is supported by European Commission by the Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems Horizon Europe research and innovation action grant (Open Music Europe 2023).

we can create new statistics, for example, from the same „microdata" that was processed to calculate the „mean age" we can create a „median age" or „top percentile age" variable. Ex ante harmonisation often aims for a different knowledge extension: we already have some processed data, and we would like to extend the inferential capacity of this dataset by collecting new data that can be harmonised with the pre-existing ones. Both forms of survey data harmonisation require a careful harmonisation of questions, answers, a processing methods.

Methodology

We decided to create a question bank application with the help of knowledge graph structured database. As opposed to a relational database, a graph database has an extensible schema. This is a natural requirement in the case of question banks, when we want to be able to survey newer and newer knowledge (competence) areas. Because we do not know what type of new questions will be added later to a question bank, and how their answers will be coded, and perhaps we also need to keep open the natural languages, we want to create thematically extendible database.

Social science researchers, but even more statistical agencies often need to harmonise surveys across different time frames, countries and languages. All European statistics that are based on questionnaires carried out among people or companies must be administered in about 25-30 languages (all official EU languages, and some member states also survey in minority languages like Catalan.) Survey harmonisation is supported by DDI, and the processing of the surveys by GSIM (Pellegrino and Grofils 2013). Neither DDI nor GSIM is fully described in the most widely used semantic interoperability standard, i.e., in RDF with an explicit ontology. A complete coverage of DDI and GSIM would allow an almost full automation of survey harmonisation. In 2024, we are far from this goal, but many important aspects of DDI and GSIM are already have semantic translations.

Our approach is taking some existing work – albeit often ongoing – in the field, and re-use these semantically rich conceptual models in a question bank application. Whenever there is no sufficient, internationally agreed consensus, we try to fill the void with a hopefully flexible, temporary solution using the Wikibase Data Model.

The Wikibase Data Model is often used as a coordinator or broker between various semantic applications due to its flexibility. For example, the Univeristy of Helsinki created a representation of the CIDOC-CM conceptual models, as international museoology standard, in Wikibase (WB CIDOC) (Kesäniemi, Koho, and Hyvönen 2022). While perhaps less efficient than a native CIDOC application, it allows easy data harmonisation via the flexible Wikibase model (the data model of the world's biggest open knowledge graph, Wikidata.) Our aim is the equivalent or near-equivalent transposition of key elements of already existing GSIM and DDI standards into a Wikibase instance (the QuestionBank.)

The description of this work is the topic of a separate paper. What we want to stress out is that our goal is to arrive at a QuestionBank application that is usable for statistical harmonisation

in the domain of music, and not a novel conceptual or information model of surveying in general. We do not aim to create an equivalent or alternative modelling of the entire DDI-Discovery vocabulary (Hartmann et al. 2024), for example, or XKOS. We want to represent the evolving standard of the DDI-Discovery and the existing XKOS standard only to the extent that we can administer our surveys and create a future-proof representation for the collected answers that allows a fully interoperable reuse capability. Describing a questions and answer options in a relatively limited universe, i.e., the socio-economic aspect of music is a relatively narrow application, although we believe that it is complex enough that our QuestionBank will be able to cope with more and more general competency requirements in the future, and can serve other creative industry research, or digital humanities research, and so on.

We aim to support this work with three R software library extensions or "packages" for a reproducible workflow:

- retroharmonize for the retrospective harmonisation of responses (Antal 2021);
- dataset for creating semantically rich representations of the response datasets (Antal 2023);
- wbdataset for connecting dataset between the R statistical system and environment and the Wikibase system (Antal 2024).

Variables

> **Note**
>
> **Variables** provide a definition of the column in a rectangular data file. Variable is a characteristic of a unit being observed. A variable might be the answer of a question, have an administrative source, or be derived from other variables. (See Q1534.)
> **RepresentedVariables** encompasse study-independent, re-usable parts of variables like occupation classification. The Representation of a variable is the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set. Representation is one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, and sex: male coded as 1). (See Q1535)

Our main focus mus be on the `RepresentedVariable` class, because these are the reusable variables that are not specific to one particular survey. We must ensure that their `Representation` and their `Question` is well-defined, and the question texts, code lists, schemas are available in all natural languages that we want to use.

While the QuestionBank should not be considered as a mini-GSIM that encompasses the statistical procedures along the questions, to allow retrospective harmonisation we must connect the Questions and Representation to some pre-existing or potential, targeted variables. A birth year may be used to infer the age of the respondent, or to create cohorts in longitudinal surveys. We do not want to foresee all potential future variable derived from the year of birth, but we must clarify that our 'In which year wereyou born?" question must result in four-digit integers (with special variables for missing or refused answers) that can be understood by a statistical application as year-precision Date/Time variable.

Most harmonised survey programs have long-established, pre-defined variables, and they often identified with a recurring numeric code or name. For example, in the Eurobarometer surveys, the D8 question .

It is a natural instinct to use D8 for both the question and the resulting variable, yet the D8 question is not the same thing as the D8 variable derived from responses to the D8 question.

In case of an ex-ante harmonisation problem, it is necessary to think about what variable we want to create as a result of the future harmonised surveying.

- The number of concert visits in a territory (an EU country, region or city) in a timeframe (usually the previous 12 month).
- etc

Representations

> **Note**
>
> The **Representation** of a variable is the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set. Representation is one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, and sex: male coded as 1). See: (Q1536)

We placed the representation base types as classes (items) into Wikibase from DDI (Data Documentation Initiative 2020) and started to create subclasses of these base types that are more specific to a questionnaire (see: representation base type (Q1384) which has part(s) of the class.)

Binary categorical variables

Often it is more efficient to establish numeric variables in a two-step process, for example, asking a person if she has a smartphone, before we try to quantify number of files or hours of various uses of the phone. These binary categorical questions also imply a numeric representation.

For example, the question "Do you own a smartphone" implies 0 smartphones in possession if answered with a `no`. In case the person owns a smartphone, an answer of 50 to the question of concert photos stored also can be processed in the following way.

0 phones ~ 0 concert photoes (0 x 0) 1 phones ~ 50 concert photos (1 x 50)

In this case, the integer coding of the a binary categorical variable makes arithmetic sense in a multiplication. We do not consider such binary questions to be numeric. Most binary questions can be used in combination with numeric variables, for example, using the similar logical arithmetic can be used to establish the highest age of female respondents in a sample of women and men (but not the average age subsetting the variable.)

Binary variables as outcome variables must be analysed with different models than continuous or categorical variables, and binary variables also behave differently as explanatory variables. Therefore, though technically speaking we could see them as categorical variables (and in some cases, as numeric ones), we clearly differentiate binary categories. Binary categorical variables are very frequently used for filtering.

> **Note**
>
> We defined the binary nominal selector (Q1399) as a subclass of the nominal representation base (Q1398): it

offers to select applicable answer options with yes/no, apply/does not apply, true/false, agree/disagree binary options.
One particular type of this representation type is the yes-no selector with decline option (Q1400). Following usual algebraic representation of logical variables, it makes sense to code yes as an integer 1, and no as integer 0, if there is no double negation in the question and the no option refers to the absence of something.

Numerically representated variables

Numerically represented variables contain only numbers. It may be necessary to contrain the accepted number range, or to create special codes for exceptions. For example, the Eurobarometer question xxxxx allows for the special answer *I'm still studying*. In this case, we end up using a mixed numeric-scale or numeric-categorical representation, which will probably yield a mixed processing (for example, a creation of a dummy variable for students, and an educational attainment proxy for those who are not studying anymore.)

Age

- Age:

Number of people

- Number of people in the household:
- Number of children in the household:

Number of activities

In the ICET model, the *E*njoyment variables tend to be activities, such as visiting concerts, listening to radio or watching movies. - Number of visits to certain cultural performances in the past 12 years.

Counting, number of aquisitions

In the ICET model, the *T*ransactions variables tend to result in the accumulation of cultural objects, although they are often just accompanying activities; for example, the purchase transaction of a festival tickets allows the enjoyment of a paid festival.

Acquisitions can be seen as activities, but they also result in the accumulation of countable things. Three repeated purchase of 2 vinyl records increases the size of the record collection by 6.

Quantities expressed in money

Quantities expressed in money, such as income, price, or spending (cost) can be expressed as numbers, however, the functional currency must be recorded in international comparison. With online surveys we must be careful if self-selecting respondents may use a different functional currency. For example, a Hungarian language self-filling questionnaire may be answered in Hungary (using HUF), in Slovakia (using EUR), and Romania (using RON).

Quantities expressed in time

Quantities expressed in time, such as minutes, hours, days may or may not be derived as intervals between points in time. Points in time variables must not be recorded as numeric variables (but using DateTime coding), however, if we ask respondents about their radio listening quantities, we will have to rely on units of measure. In the Hungary CAP survey, we used a simple model to establish annual notional enjoyment hours.

Daily frequency (250) x 2 hours of music ~ 500 hours per year. Monthly frequency (12) x 60 minutes of music ~ 12 hours per year.

While we are aware of the biases of such self-estimated time intervals, we believe that while the time intervals are biased, the ratios of time intervals are unbiased or much less biased estimators.

Geographical distances

Geographical distances can be measured with time length of walking, biking, driving, or estimated meter, kilometres, or miles travelled. When we have relatively exact locations (respondent lives in Debrecen, visits the Budapest Arena) we may get a better estimate of the travel distance (in time or in space) by using well-established proximate distances of any point of Debrecen to the Budapest Arena than asking people directly.

Geographical representations

- Country of residence:
- Country of birth:
- Sub-national region of residence:
- Location of the last visited concert:
- Postal code districts: special numeric coding for representing relatively small, sub-national areas.

- Important venues: we may ask about visiting (binary or frequency) of some nationally important venues, such as the National Opera or the Budapest Arena, such venues also imply visiting a specific location.

Scale representation

- Subjective economic position:
- Simplified educational level:

Psychometric variables

In the Hungarian and Slovak CAP surveys we have experimented with the use of simplified psychometric variables.

Questions

> **Note**
>
> A `Question` is designed to get information upon a subject, or sequence of subjects, from a respondent. (See Q1295.)

We placed several questions into the Eurobarometer trend questions (Q1310) demonstration question bank.

Table 1: Our representation of the disco:Question class

Property of disco:Question	Wikibase model
responseDomain	added as a statement, must be an instance or subclass of a representation (Q1536)
questionText	description (long label) of the Wikibase item.
(identifier)	the QID of the question
rdfs:label	label (short label) of the Wikibase item

> **Caution**
>
> Our model still needs refinement! The example below is not the final model.

Consider the the following three entities (items), which connect a question with a representation and a variable.

- Age education variable (Eurobarometer-GESIS) (Q1530)

  a sociodemographic categorical variable derived from the question "How old were you when you stopped full-time education?"

- age when finished education (Eurobarometer) (Q1405)

  How old were you when you stopped full-time education?

- age when finished education coding (Q1411)

  If "STILL STUDYING", code '00'; if "NO EDUCATION" code '01'; if "REFUSAL" code '98'; if decline "DK" code '99', othervise answers are coded as numbers.

The last bit needs to be thought through before uploading much data!

Socio-demographic questions

Table 2: Eurobarometer socio-demographic questions

Question	Variable
How old were you when you stopped full-time education? (Q1405) type of community (Eurobarometer) (Q1294)	Age education variable (Eurobarometer-GESIS) (Q1530)

References

Antal, Daniel. 2021. "retroharmonize Ex Post Survey Data Harmonization." The Comprehensive R Archive Network. https://doi.org/10.5281/zenodo.5781724.

———. 2023. "Dataset Create Interoperable and Well-Documented Data Frames." The Comprehensive R Archive Network. https://doi.org/10.5281/zenodo.10304055.

———. 2024. "Wbdataset Making Datasets Truly Interoperable and Reusable in r with Wikibase." Zenodo. https://doi.org/10.5281/zenodo.10304055.

Data Documentation Initiative. 2020. "DDI Lifecycle (3.3) Documentation." https://ddi-lifecycle-documentation.readthedocs.io/en/latest/index.html.

Hartmann, Thomas, Sarven Capadisli, Franck Cotton, Richard Cyganiak, Arofan Gregory, Benedikt Kämpgen, Olof Olsson, Heiko Paulheim, Joachim Wackerow, and Benjamin Zapilko. 2024. "DDI-RDF Discovery Vocabulary. A Vocabulary for Publishing Metadata about Data Sets (Research and Survey Data) into the Web of Linked Data." Edited by Thomas Hartmann, Richard Cyganiak, Joachim Wackerow, and Benjamin Zapilko. W3C. https://rdf-vocabulary.ddialliance.org/discovery.html.

Kesäniemi, Joonas, Mikko Koho, and Eero Hyvönen. 2022. "Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC CRM." In *New Trends in Database and Information Systems*, edited by Tania Chiusano Silviaand Cerquitelli, Robert Wrembel, Kjetil Nørvåg, Barbara Catania, Genoveva Vargas-Solar, and Ester Zumpano, 542–49. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-15743-1_49.

Open Music Europe. 2023. "Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems." https://doi.org/10.3030/101095295.

Pellegrino, Marco, and Denis Grofils. 2013. "DDI-SDMX Integration and Implementation. Working Paper." United Nations Economic Commission for Europe. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2013/WP5.pdf.