A Documentation Model for Comparative Research Based on Harmonization Strategies

John Kallas & Apostolos Linardis[1]

Abstract
This paper deals with studies of fixed design, where comparability is feasible either for culture or for time and culture. For the time dimension, studies are divided into cross-sectional and longitudinal, and for the cultural dimension they are divided into mono-cultural and cross-cultural. Since modern societies are mainly organised into nation-states, cross-cultural studies are carried out more often on country level. However this does not necessarily mean that the organisation of cross-cultural studies in the same country is not possible. Comparative cross-cultural studies follow strategies of data harmonization such as the «ex ante input harmonization», the «ex ante output harmonization», and the «ex post harmonization», as well as mixed strategies. These strategies of data harmonization are complex procedures for which success or failure is reflected in the final study product.

In this paper, a documentation model is proposed for both longitudinal and cross-cultural studies, for which documentation is particularly difficult. All the remaining study types can be considered sub-cases of the longitudinal and cross-cultural study and consequently are covered by the proposed model. Three documentation models are proposed according to the different harmonization strategies examined. Finally, the different documentation models are integrated into one.

All the models examined are drawn as entity relationship diagrams based on the harmonization methods that govern comparative research.

Keywords: research documentation, data modelling, comparative research, harmonization strategies.

Introduction
Comparative cross-cultural studies follow different strategies of data harmonization such as the «ex ante input harmonization», the «ex ante output harmonization», and the «ex post harmonization», as well as mixed strategies. These strategies are complex procedures for which success or failure is reflected in the final study product.

In this paper, a documentation model is proposed for both longitudinal and cross-cultural studies. Three documentation models are proposed according to the different harmonization strategies examined. Finally, the different documentation models are integrated into one.

1. Cross–national studies as a sub-case of cross-cultural studies
The original pattern of comparative research brings together at least two social formations. Some researchers identify cross-national with cross-cultural studies. Thus, for Hantrais (1995), comparative research is a research pattern of the social sciences that aims to conduct comparisons between representations that result from two or more social formations. Globalisation and the revolution in communication technology on one side and the developments in Europe and the course of its unification on the other side, bring in question our current understanding of a «nation–state» as well as the convention to consider comparative research simply as cross-national research.

Globalisation began in the economic sector and next swept across sectors of policy, culture, and knowledge production. The consequence of this evolution was the gradual delimitation of relations, and of the role of the nation-state (Albrow 1998). This change became evident in social science literature as the loss of territoriality followed by the reduction in the sovereignty of nation-states and denationalization (Zurn 1998).

Despite the controversial nature of the «nation-state», most cross-cultural studies are still organised as cross-national ones. However, in order to also cover the case of cross-cultural studies that are not cross-national, we will use the term «cultures» instead of the term «countries» which is commonly used in cross-national studies.

Moreover, as previously stated, cultural discrepancies may exist in the same «nation–state» either on a regional or local level. A basic cultural difference is language. Countries such as Belgium, Finland, and Luxembourg are obligated to carry out any national study in multiple languages. The action plan of the European Science Foundation reports explicitly: «Translations should be made into any language

which is used as a first language by five percent or more of a country's population» (1999, 10).

This leads us to consider cross-cultural research to have a wider nature than cross-national research and cross-national research as a sub-case of cross-cultural research with a determined culture, namely the nation–state.

2. Formal description of the data element in cross-cultural research

In empirical fixed-design studies (Robson 2007), data production is organised based on a data schema. This data schema is constructed on the basis of statistical ontology, which predicts that the examined population is constructed of similar units of observation, each of which is described by concrete, distinguishable attributes, each of which is represented by a variable. The adoption of statistical ontology as an organisational model of the data schema of fixed-design studies has two basic advantages. The first advantage is that it allows the application of statistics as a method of data analysis. The second advantage is that it allows for the development of a documentation initiative for studies of fixed design. Each population attribute is formally described by a data element that is defined by one concept and by one pattern of value determination (Kallas and Linardis 2009). Each concept is defined by one term and one definition. Each value determination pattern is defined as a) a mono-dimensional classification, b) a number that results from the direct measurement of a concept, or c) text.

The unit of observation, as it is introduced by statistical ontology, is a mathematical schema that does not always correspond to real objects of observation. That is, to social objects that are presented in social practices independently from the observer. The relation of a unit of observation with a real object of observation occurs in the context of each concrete study. In certain cases where the examined social phenomenon consists of the relationships between more than one object of observation (for example in the case of a household in which we usually have at least two objects: the household and the members of the household), the unit of observation does not correspond to real objects of observation but simply represents the total set of all attributes of individual objects (Kallas 2005).

In cross-cultural studies where partial recordings describe different societies (the objects of observation based on which the formal descriptions of social phenomena are constructed) it is possible to differentiate from society to society and consequently from recording to recording. This difference means that either the corresponding objects between two recordings cannot be described by the same data elements, or that certain data elements are not precisely the same.

Consequently, in cross-cultural research, a data element should be described in reference to the object of observation that it describes. The object of observation in each study is related to one unit of observation, which should be documented based on the following:

a) the definition of objects of observation that make up the unit of observation;

b) the social system in the context of which the objects of observation are constructed. Each object of observation is notionally defined in the context of a concrete social system. When it is used in the context of another social system then it is differentiated notionally, and consequently it is described via a different pattern. These differences also concern the data elements whereby the object is defined;

c) the territory where the recording is organized.

The universe is defined by a) a territory, b) a unit of observation and, c) other partial determinations (such as age, sex, marital status, income etc.). The universe refers to both the data elements as well as to the objects of observation. Vice versa, each data element or object of observation refers to a universe. A documentation schema of a data element suitable for the cross-cultural research is shown in figure 1 (see next page).

It is possible to define a territory by a multilingual, controlled vocabulary that includes country and region names (for example the Nomenclature of Territorial Units for Statistics or NUTS classification) as well as other predefined values such as international, European, etc. It is also possible to define the object of observation and social system by a controlled vocabulary. This standardisation further helps in comparative research since the objects of observation are often used by researchers as basic search and comparability criteria.

The documentation schema of a data element includes three basic structural components: the concept, the universe, and the classification. It is possible to reuse the same data element for the documentation of one or more studies. It is also probable for some other data elements to be identified partly by reusing a subset of the structural components that compose a data element. In addition, components such as the universe or category schema can also be reused for the determination of other entities. For example, the category schema can be reused either for the determination of a classification, for the determination of a question, or for the determination of a variable. Additionally, the universe can also be reused for the determination of study «wave instance». However, the universe of each data element usually constitutes a subset of the general universe of a wave instance. In most cases, both universes are similarly identified at the territory and unit of observation level but differ in their other determiners.

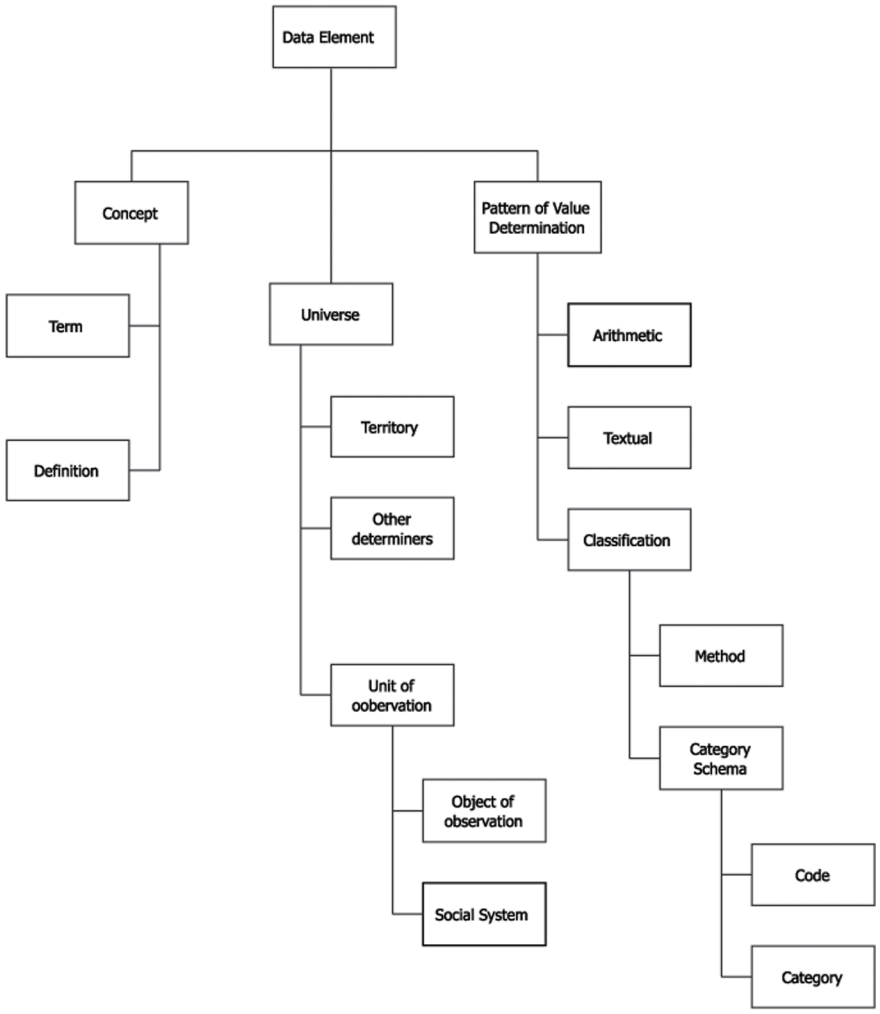

Figure 1: Documentation schema of a data element

The data element, the concept, the universe, and the classification are study components for which even small changes in the content should be documented since they may affect the overall comparability. This is the reason why they are defined as versionable objects. These objects are identified by one code but also by a version number (complex key). Both major and minor changes in the lower level objects simultaneously alter the version of parent objects. The version change documentation should include the following: a) the new version date, b)who made the change, c) why there was a change, so that the users can comprehend if the version change influences the analysis of data (DDI Alliance).

3. Approaches in data harmonization

Harmonized data can be achieved either by using strategies for collecting harmonized data from the beginning (in the study design) or by using harmonization strategies for existing data (Granda, Hadorn and Wolf 2008).

3.1 Strategies for the production of harmonized data in cross-cultural studies

The following is a short description of strategies followed for the production of harmonized data in cross-cultural studies.

- Ex ante input harmonization

Ex ante input harmonization means that the institutions that participate in the study have agreed on common concepts, common measurement patterns of the concepts and also on common questions based on a common source questionnaire. The ex ante input harmonization is used mostly in cross-national studies but is also applied in studies that are conducted in the same country. For example, a question relative to the underground in Greece, such as «How many times do you use the underground per week? » is asked only to Athenians but not to all Greeks. Consequently, an agreement is required for the concepts, the measurement patterns, and the questions even in the same country in order for the questions to have meaning for all participants. In cross-cultural studies where ex ante input is applied, no country-specific variations are allowed except the ones that are absolutely essential such as the language used in the questionnaires (Ehling 2003).

- Ex ante output harmonization

In ex ante output harmonization the institutions that participate in the study have agreed on common concepts and common measurement patterns. The objective is fixed and the choice of suitable questions is left to participating research groups who adapt the questions to the cultural particularities of the universe that they study. Each research group determines its own concepts and measurement patterns; however, they must correspond to the common concept via transformation routines. For example, let us assume that in a cross-national study the participants are asked to indicate their highest level of education. A

common measurement pattern for education level is the use of an international classification such as the International Standard Classification of Education (ISCED). The measurement of education level via ISCED may serve international needs but not national ones, since a country would serve itself more if its national data were detailed enough. Another reason to follow this strategy is that the same data can be collected simultaneously for different studies. More concretely Lene Mejer reports:

«Output harmonization means to give a common internationally agreed definition for a variable and then leave to each single Member State to decide on its implementation. Each Member State decides what is the best national source for the variable (for example from already-existing surveys and / or registers) » (2003, 69).

This strategy is used mostly for cross-national studies; however, it can be used as methodology in cross-cultural studies.

- Mixed Strategy

Some studies, while they follow the strategy of ex ante input harmonization (such as the European Social Survey[2] or ESS and the International Social Survey Programme[3] or ISSP), in certain selected data elements they apply the ex ante output harmonization strategy. For example the highest education level in ESS is ex ante output harmonized while the study is ex ante input harmonized. In this case, the harmonization strategy should be placed at data element level per wave and not at study level. There is another case of «mixed strategies» where the literal question is agreed upon by the harmonization committee but the category schema is fixed by each country separately.

3.2 Harmonization strategy for existing data: ex post harmonization

Ex post harmonization is a harmonization strategy where the total study results from already-existing studies. In ex post harmonization, the institutions that participate in the study agree on common concepts, on common measurement patterns, and on common universes (common data elements). They also agree on already-existing studies that have to be ex post harmonized using transformation routines. The achievement of harmonized data via this process is not guaranteed, even if it has been optimally designed, because of the diversity of concepts and measurement patterns in existing studies. Since no new questions are created, the basic structural elements of these studies are the data elements. New data elements are created that reference already-existing ones. For the implementation of such studies (that resemble research programs more than studies) transformation routines are required, which are written with statistical software. The difficulty of implementing studies following the strategy of ex post harmonization lies in the localisation of common concepts and measurement patterns between the universes. A very useful tool for such studies would be a bank of concepts, classifications, and universes for the localisation of similar data elements.

4. Data archives and the documentation process: a comparative perspective

In recent years, new organizations have been created in Europe called data archives (DA). DA deal with the accumulation, documentation, and dissemination of data. These organizations support secondary analysis and comparative research, and act as mediators between the producers and analysts. The European council is called the Council of European Social Science Data Archives (CESSDA). Each DA must document its own studies based on a common strategy that ensures the following:

- Reuse of common structural study components of a simple study in the same DA

Each study component can be constructed from other structural components. For example, study components such as classification, question, and variable use some common structural components such as codes and categories. Often the category schema of classification, question, and variable coincide. For example, the codes and categories that are used for ISCED classification (for the corresponding question but also for the corresponding variable in a statistical data file) may coincide. In this case the study components' common structural components should be imported just once and then reused (DDI Alliance) even in the case of a simple study that is neither longitudinal, nor cross-cultural. Each person who documents a study should follow these rules so that double entries are avoided. To aid in this laborious documentation work, certain processes for localisation of common structural components can be automated.

- Comparability of a longitudinal study in the same DA.

- The comparability of a longitudinal study in the same DA lies in the reuse of study components between waves. Components such as concepts, classifications, universes, questions, and variables are principal components for comparability between waves of longitudinal studies and they should be reused in the various waves. Consequently, each DA should maintain local banks of all these study components.

- Comparability of different studies in the same DA

Comparability of different studies is also based on the reusability of the same principal study components

between the various studies, as in a longitudinal study.

- Comparability of a cross-cultural study in the same DA

The proper documentation of a cross-cultural study involves all the participating organizations and it differs depending on the harmonization strategy that has been followed. The documentation of a cross-cultural study based on the harmonization strategy followed is developed analytically in section five. While the documentation of a cross-cultural study often occurs in different DA's, in this work we will deal with the documentation of a cross-cultural study in the same DA (or data-metadata repository).

- Study comparability in different DA's

Study comparability in different DA's is a very critical process for wider comparative research but this will be analyzed in a later work.

Summarizing the above, it is immediately evident that the documentation procedure is a difficult and laborious process. On the other hand, the result of this procedure will be useful for the wider research community, particularly for those researchers who want to carry out comparative research and secondary analysis. The comparative documentation further strengthens the role of DA's. The documentation process is best carried out in collaboration with the primary data producers as well as with the statistical institutes.

5. The documentation of a cross-cultural, longitudinal study

The documentation process is rendered particularly difficult and laborious in the case of cross-cultural, longitudinal research. The collaborating institutions should follow the documentation of the coordinating institution, since the resulting documentation will be based on common agreed concepts, measurement patterns, questions, and universes. The documentation completed by the coordinating institution should not be changed by the participating institutions. The documentation language of the coordinator is the common agreed language (usually English).

In the cases that follow, the model is presented first and then a description of how the model should be used by participants based on the agreed-upon harmonization strategy. It takes into consideration the most complex study type, the cross-cultural, longitudinal study, since all other studies can be documented based on this. The diachronism relies on the reuse of study components for each wave. The multiculturalism lies in the creation of references between the source and universe study components.

It should be noted that the models that follow concern the most complex study type – the cross-cultural, longitudinal study – but they document just one study. Another limitation is that the documentation takes place in the same DA and not in distributed documentation systems.

Below is a short description of the main entities used in the models:

- Study: the entity that is used to store the general study information such as title, more general objectives, summary etc. The main purpose of adopting such an entity, beyond the storage of general information, is that it aims to unify all study waves. The study level documentation is completed by the coordinating institution in the common agreed language. Translation into other languages occurs only for dissemination reasons. This entity is not reusable but can be referenced by other studies or by other study components.

- Wave: a longitudinal, cross-cultural study takes place in many time and universe instances. The time instance of a study is called a wave while the universe instance of a study wave is called a wave instance. While documenting, but also while a study is conducted, it is common practice to establish the time and, for each time period, to receive snapshots for the various universes that participate in the study. In addition, at this level, the general wave title, any special objectives per-wave, and the total duration of the study wave are all recorded. Information such as the universes or institutions that participate in the study may be recovered automatically from the wave instance documentation level so that no differences in the aggregated fields of wave level exist. The most crucial documentation at wave level has to do with the determination of the study harmonization strategy. It is also crucial this be selected from a controlled vocabulary where the user chooses between the following options: a) ex ante input harmonization, b) ex ante output harmonization, c) ex post harmonization, or d) mixed strategies. The harmonization strategy is determined at wave level, not at study level, because it is possible (although rare) that the harmonization strategy may change from wave to wave. The wave entity is also used for the grouping of source data elements, of common source questionnaires (when they exist), and of the harmonized statistical data

  files. The documentation at wave level should be completed by the coordinating institution in the common agreed language. Translation into other languages is done only for dissemination reasons. The wave entity cannot be reused but can be referenced by other study components..

- Wave instance: the entity that is used for storing information concerning wave snapshots per universe. The documentation of wave instance level is completed by all the research groups that participate in the study, in the languages that have been decided per group but also in the common agreed language for dissemination purposes. The wave instance includes extensive information such as universe, sampling methods, participating institutions by role (local coordinator, financiers, data producers, organizations responsible for data dissemination), researchers, sampling frame, data collection method, time of data collection, and description of weights (accompanied by weighting methodology). The wave instance is not a reusable object but can be referenced by other study components.

- Source data element and universe-specific[4] data element: entities used to store information concerning the data elements that were introduced in section two. The implementation of these two entities in a database does not necessarily require the creation of two tables for the two types of data elements; however, both data elements are presented as separate entities in the entity relationship diagrams in order that the required relationships are evident. The same holds for the questionnaires, the questions, the data files, and the variables. The data element and its structural components are reusable entities for different studies or study waves.

- Source questionnaire and universe-specific questionnaire: entities used to store general information concerning the questionnaire such as the number of questions, type of questionnaire (standardized versus non-standardized), abstract, and link to the questionnaire file. This is also a grouping entity for questions. Questionnaires are not reusable entities but have to be defined again in each wave or wave instance.

- Source question and universe-specific question: according to Kallas and Linardis (2009), the questions are composed of some or all structural elements presented in figure two. The question is also a reusable object.

- Harmonized data file and universe-specific data file: entities used to store general information about the statistical data files such as the number of variables, the number of cases, and likely a link to the statistical data file. Data files are also used as grouping entities for the variables. The data files are not reusable entities and have to be defined again for each wave or wave instance.

- Harmonized variable and universe-specific variable: the variables consist of structural elements such as name, description, type, measurement level, and category schema. The variable entity, as it is described here, consists only of metadata and not of data. The same variable may have a number of data depictions but in different statistical data files. The variable is a reusable entity.

- Finally, the transformation routine is the process that describes the necessary transformations of the universe-specific data element to source data element.

5.1. Case 1: study documentation following ex ante input harmonization strategy

The documentation process for ex ante input harmonization is portrayed in figure 3 (on next page). The left parallelogram portrays the documentation that should be completed by the coordinating institution while the right one portrays the documentation that should be completed by the participating institutions.

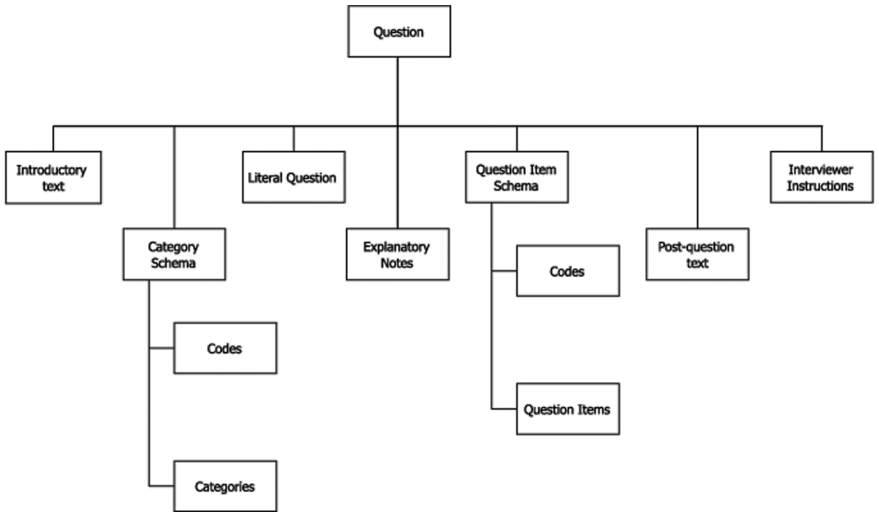

Figure 2; : Documentation schema of a question

The documentation may be published in intermediary stages (indicated below). The basic principle of ex ante input harmonization is the standardization of data elements and questions between the research groups. The stages of documentation are as follows:

Stage 1: Documentation of the general context of the study and wave, and documentation of data elements (documentation provided by coordinating institution).

1. Documentation of the general context of the study: Documentation is completed by the coordinator at the beginning of the study.

2. Documentation of the general context of the study wave: Study waves have to reference the corresponding study.

3. Documentation of common data elements (common concepts, measurement patterns, and universes): Data elements should reference the study wave and not the study because the data elements may differ from wave to wave. For example in the ESS, there are some data elements that are used in all study waves while others are added or removed periodically.

Fundamental practices to ensure comparability: dissemination of stage one documentation to the participating institutions.

Stage 2: Translation of the context of study and study wave; translation of the data elements; determination of each wave instance (documentation provided by participating institutions).

1. After the dissemination of stage one documentation to the participating institutions, each participating institution should translate the general context of the study and of the study wave, and the data elements, for dissemination reasons.

2 Each participating institution should then document the wave instance based on the universe it represents. Each wave instance has to reference the corresponding wave.

(First intermediate phase of study publication)

Stage 3: Documentation of source questionnaire/s (documentation provided by coordinating institution).

1. Documentation of source questionnaire/s: Each questionnaire should reference the corresponding wave.

2. Documentation of source questions: The reference between source questions and the corresponding data elements as well as between source questions and the source questionnaire is required.

Fundamental practices to ensure comparability: a) dissemination of stage three documentation to the participating institutions, b) creation of a statistical file template with common variable names (based on the data elements) for all participating institutions, c) dissemination of the template to the participating institutions.

Stage 4: Translation of source questionnaire/s, leading to universe-specific questionnaire/s (documentation provided by participating institutions).

1. Documentation of universe-specific questionnaire/s: The reference between the universe-specific questionnaire and the corresponding wave instance is required.

2. Documentation of universe-specific questions: The questions specific to each universe are created via translation of the source questions in language or languages decided by each research group. The reference between universe-specific questions and source questions as well as with the corresponding universe-specific questionnaire is required.

Fundamental practices to ensure comparability: a) each institution conducts the research, collects the data, and submits the statistical data files to the coordinator, according to the template already sent by the coordinator, b) at the same time, each institution preserves the data files for the stage six documentation procedure.

(Second intermediate phase of study publication)

Stage 5: Documentation of harmonized statistical data file/s (documentation provided by coordinating institution).

1. Documentation of harmonized statistical data files that have come from the merging of universe-specific data files: The reference between harmonized statistical data file/s and wave is required.

2. Documentation of harmonized variables: The reference between harmonized variables, the harmonized statistical file, and the corresponding source questions is required.

Fundamental practices to ensure comparability: dissemination of stage five documentation to the participating institutions.

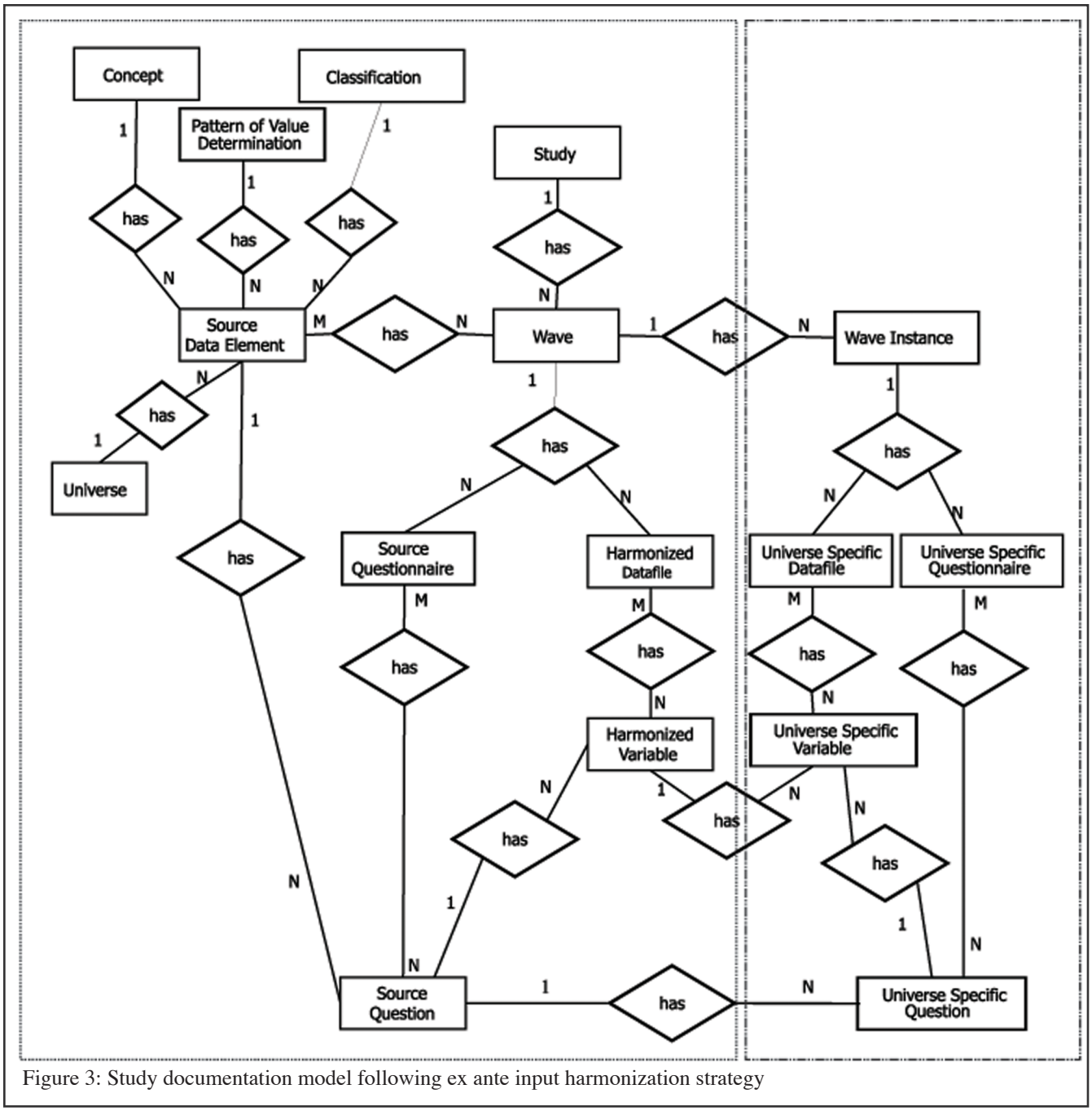

Figure 3: Study documentation model following ex ante input harmonization strategy

Stage 6: Documentation of universe-specific data files (documentation provided by participating institutions).

1. Documentation of universe-specific statistical data files: The reference between universe specific statistical data files and the corresponding wave instances is required.

2. Documentation of universe-specific variables: The reference between universe-specific variables and the statistical data file they belong to, universe-specific variables and the corresponding universe-specific questions, as well as reference between universe-specific and harmonized variables is required.

(Final phase of study publication)

5.2. Case 2: study documentation following ex ante output harmonization strategy.

A basic difference between ex ante input harmonization strategy and ex ante output harmonization strategy is that the second presupposes the determination of universe-specific data elements by the participating institutions. Consequently, the participating institutions should document the universe-specific data elements and reference them to the common agreed data elements via transformation routines. Also, there is no source question, just a source data element. On the other hand, there are universe-specific questions but these are considered mostly as additional documentation of universe-specific data elements not as fundamental structural study components. The documentation process for ex ante output harmonization strategy is portrayed in figure 4. For simplistic reasons we have not drawn the structural elements of the data element again (concept, measurement pattern, and universe). The left parallelogram portrays the documentation that should be completed by the coordinating institution while the right one portrays the documentation that should be completed by the participating institutions.

Following this strategy, there are five study documentation stages instead of six. This occurs because stage three of ex ante input harmonization does not make sense here since there are no source questionnaires or source questions. The

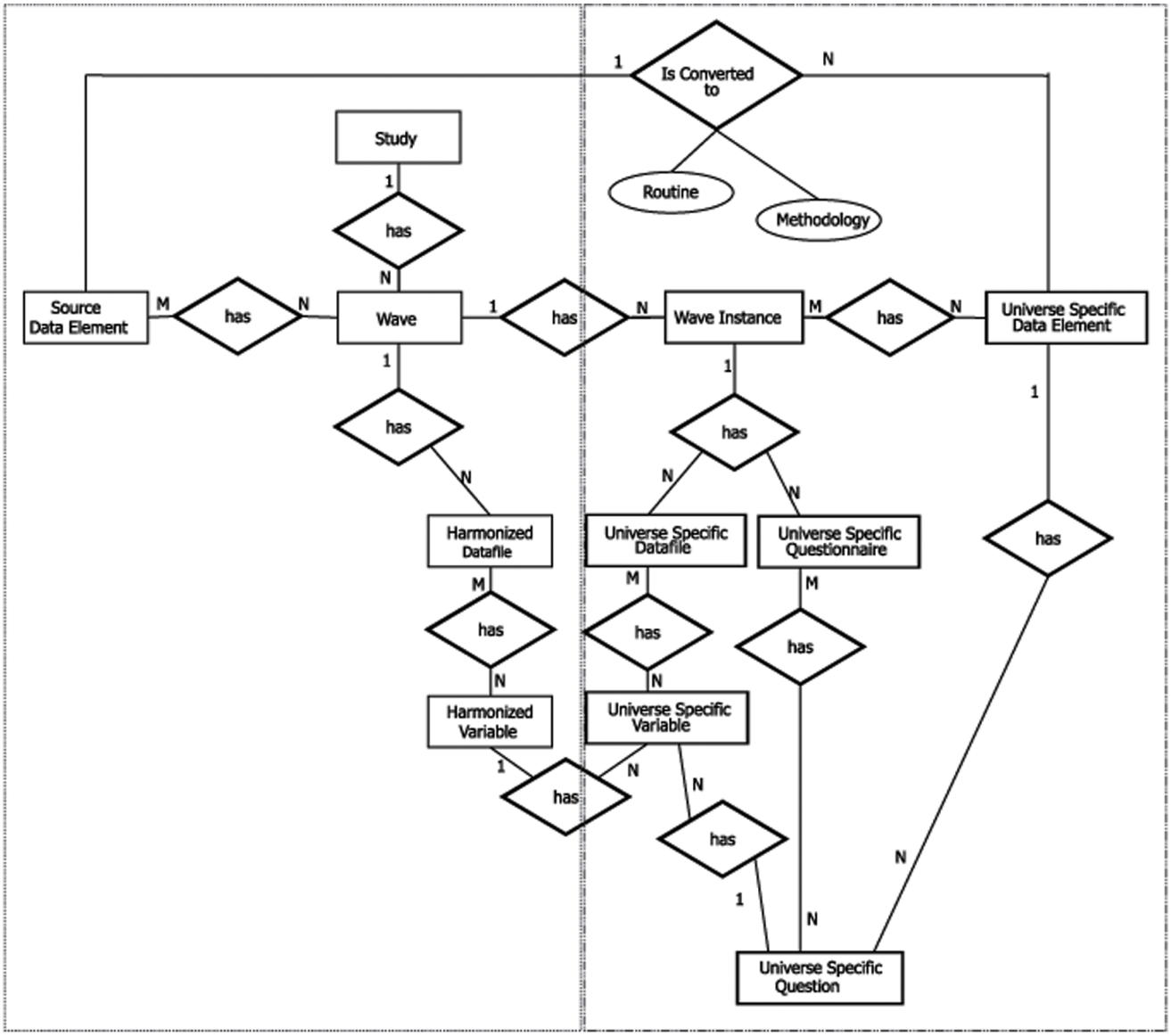

Figure 4: Study documentation model following ex ante output harmonization strategy

differences between the documentation procedures of the two strategies are summarised as follows:

- Stage 2. This stage includes two additional documentation actions to be performed by the participating institutions: 2.3) documentation of universe-specific data elements; and 2.4) documentation of transformation routines of source to universe-specific data elements.

- Stage 3. As was mentioned before, stage three does not exist. Nevertheless, the coordinating institution can establish some fundamental practices to ensure comparability such as: a) the creation of a statistical file template with common variable names for all participating institutions, based on the data elements; and b) the dissemination of the template to the participating institutions.

- Stage 4. Phase 4.2 is different because there are no source questions. Consequently, the universe-specific questions are developed from scratch instead of being produced as translations of the source questions. Reference between universe-specific questions and questionnaire as well as between universe-specific questions and data elements is required.

- Stage 5. Phase 5.2 is different because there is no reference between harmonized variables and source questions. Reference between harmonized variables and harmonized statistical file is required.

5.3. Case 3: study documentation following ex post harmonization strategy

The basic difference between this strategy and ex ante output harmonization lies in its relationship to the common concept. In ex ante output harmonization, its relationship to the common concept is guaranteed because the researchers design the wave instances from scratch, keeping in mind the common data elements. In ex post harmonization, the individual studies have been designed autonomously by the researchers without adhering to a common concept. Thus, the relationship to the common concept is not guaranteed. On the other hand, the two harmonization strategies have a lot of similarities related to methodological issues. The documentation process is similar to the one that was described based on figure 4. The basic difference is that in ex post harmonization, the new study derives from already-existing studies. Consequently, it should initially be documented using the already-existing study waves or wave instances from which the new study derives. It would provide great relief from the excessive documenting load for researchers working on an ex post harmonized study if documentation of existing study waves or wave instances was already available.

Let us assume that a new study is designed following the ex post harmonization strategy. The new study concerns attitudes for a set of countries (two of which are: Cyprus and Russia), for the time period 2004-2005. The coordinating committee decides to harmonize ex post the second wave of the ESS. According to Jowell et al. (2007), Cyprus and Russia did not participate in the second wave of the ESS. Nevertheless, the coordinating committee is aware of the existence of other national attitude studies for Cyprus and Russia during 2004-2005 and decides to harmonize them ex post. At the same time, the coordinating committee has to decide on the common data elements of the new study. After the common data elements have been defined, the data elements of the existing surveys have to be transformed via routines to the common ones. It is common for different groups to undertake the transformations of different studies. In our example, three groups will undertake the burden of transformations of source data elements to the common agreed data elements: one group for the ESS, one for the Cyprian study, and a group for the Russian one. The documentation process of the three groups includes: a) the documentation of each new wave instance; b) the reference of already-documented data elements (ideally) to the wave instance; or c) the documentation from scratch of all previously conducted studies in the research program, if their documentation does not exist; and d) the application of transformation routines to universe-specific data elements and source data elements.

The documentation process in ex post harmonization is portrayed in figure 5. The left parallelogram portrays the documentation that should be completed by the coordinating institution of the research project, while the right one portrays the documentation that should be completed by other institutions that participate in the research project. These institutions will have undertaken the documentation of concrete wave instances from already-existing studies.

Figure 5 differs from figure 4 in the following ways:

- Each new study wave can be designed based on existing study waves and/or existing wave instances. Consequently, suitable documentation at wave level is required.

- The relationship between source data element and universe-specific data element is a “many to many” relationship, since the same universe-specific data element may correspond to more than one source data element in the same system. For example, a universe-specific data element may correspond both to the source data element of the

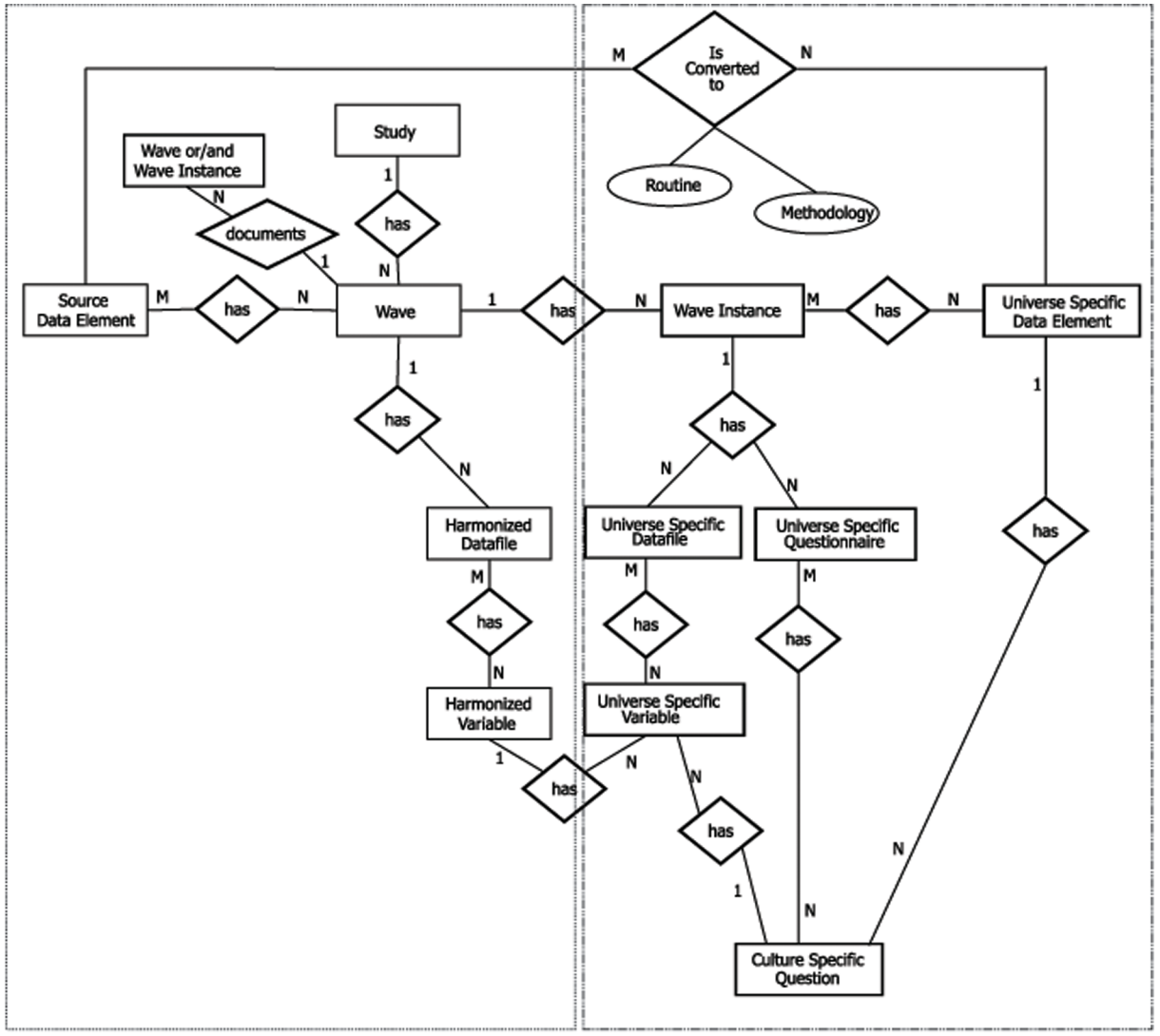

Figure 5: Study documentation model following ex post harmonization strategy

initial study and to a new source data element that was created during the design of a study following ex post harmonization.

5.4. Common documentation model for all harmonization strategies

As already mentioned, the harmonization strategy followed should be determined at wave level. In the case where the strategy followed is ex ante input harmonization, the documentation model is the one in figure 3. The documentation models of ex ante output harmonization and ex post harmonization are depicted in figure 4 and figure 5, respectively. In the case where the harmonization method is determined as «mixed strategies», the harmonization strategy has to be determined for each data element per-wave. The entity relationship diagrams 3, 4 and 5 are unified in figure 6. In using this model, the documentation process will differ depending on the choice of harmonization strategy.

The documentation model in figure 6 also serves the needs for the documentation of different studies based on a comparative perspective. This is feasible because entities such as concept, classification, universe, data element, question, and variable, that constitute the basic study components for comparative research, are reusable entities for all studies. The reusability of these study components, in a documentation system of a specialized architecture, aims at comparative documentation between different studies. Another useful outcome of such a documentation

model is that it is feasible for a researcher to locate universe-specific study components derived from source study components.

When they are referenced, the study components referred to above can never be deleted or changed. These components are identified by Persistent Identifiers (PIDs). If one of these components has to change then a new version of this component has to be created.

Multilingualism of study components is applied in two cases: a) when a component is translated by an institution in order for its translation to be an "active component" of the study (for example translation of the source questions to universe-specific questions); and b) just for dissemination reasons (for example translation of a study's abstract). The reasons for translation of a study component should be declared in the documentation. In the first case, both major and minor changes may lead to version change of the study component, not so in the second one.

6. Conclusions

The general documentation process in a DA or in a metadata and data repository is based mostly on ex post harmonization procedures. The institutions that document a longitudinal, cross-cultural study should do so based on already-existing documentation. In the case of longitudinal studies, this should be done by repeating study components from other waves, and in the case of cross-cultural studies, by referencing source and universe-specific objects.

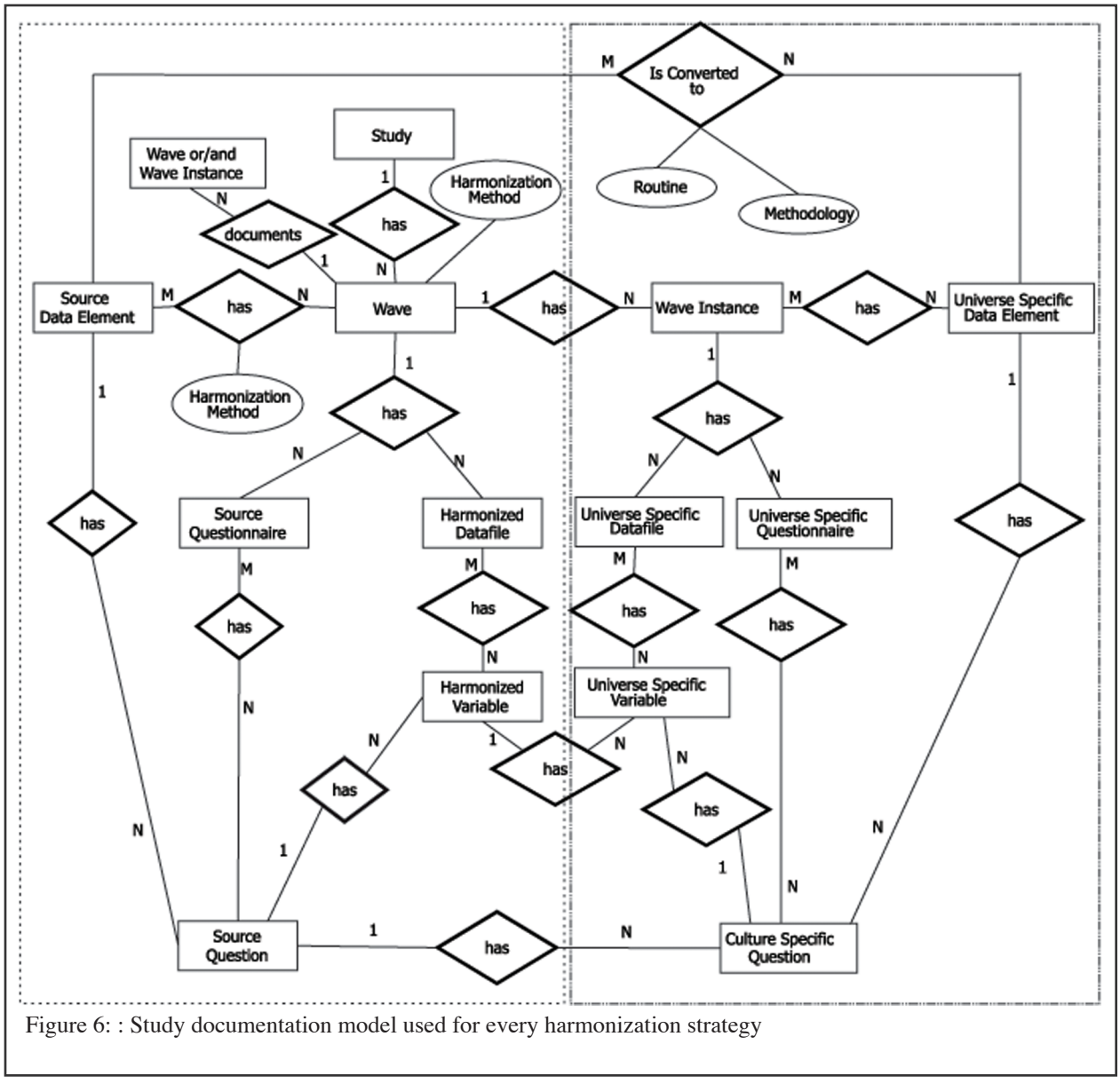

Figure 6: : Study documentation model used for every harmonization strategy

Consequently, the documentation procedure not only involves the typical description of the components derived in the context of a single research project, but also all of the a posteriori references (additional documentation) between independently-designed studies.

The proposed study documentation procedure is laborious for the researchers that are making the documentation. Additionally, s/he should know the specifics of each study, which presupposes a close collaboration with the primary investigators. Moreover, the "golden super rule" of METANET research project states: «Metadata are as important as data, and metadata need as much work as data» (Sundgren, 2003, 129). The study documentation with suitable metadata can provide particular advantages in the identification of «equivalent» or «equal» study components either for longitudinal or cross-cultural studies, or even studies with similar subjects.

Documenting studies based on the proposed documentation model, allows the researcher to:

- search and locate questions that use either common concepts, common classifications, questions that are addressed in common universes, or even questions that use common data elements;
- search all the variables that are derived from the same questions;
- locate data element transformation routines, so that the sequence from one data element to the other is clear;
- locate all the translations of a source question, likely accompanied by qualitative criteria such as validity and reliability but also non-response rates (Sarris et al. 2007);
- locate data according to the following criteria: a) the concepts that the data imply, b) the universes the data refer to, c) the time period the data refer to or the time the survey was conducted, d) concrete classifications based on which the data have been produced.

This model can also be extended and modified so that it may cover the documentation of comparative research in distributed environments. In this case, not only the model would be of particular interest, but also the flexible and functional architecture of the overall documentation system. The extension of the model in distributed environments as well as the architecture of such a system will be analyzed in a future paper.

References

Albrow, Martin, and Raimund Fellinger. 1998. Abschied vom Nationalstaat: Staat und Gesellschaft im globalen Zeitalter. Frankfurt am Main: Suhrkamp.

DDI Alliance. DDI 3.0. http://www.ddialliance.org/ddi3/index.html

Ehling, Manfred. 2003. Harmonizing data in official statistics: Development, procedures, and data quality. In Advances in cross-national comparison: a European working group for demographic and socioeconomic variables, ed. Jurgen H.P. Hoffmeyer–Zlotnik and Christof Wolf, 17-31. New York: Kluwer Academic / Plenum Publishers.

European Science Foundation (ESF). 1999. Blueprint for a European Social Survey. Strasbourg: Author.

Granda, Peter, Rheto Hadorn, and Christof Wolf. Harmonizing survey data. Paper presented at the International Conference on Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC), June 25-28 in Berlin Germany.

Hantrais, Linda. 1995. Comparative research methods. Social Research Update 13, (Summer). http://sru.soc.surrey.ac.uk/SRU13.html

Jowell, Roger, Max Kaase, Rory Fitzgerald, and Gillian Eva. 2007. The European Social Survey as a measurement model. In Measuring attitudes cross–nationally: Lessons from the European Social Survey, ed. Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 1-31. Los Angeles: Sage.

Kallas, John and Apostolos Linardis. 2009. Questionnaire documentation model on the needs of comparative research. Paper under review for publication.

Kallas, John. 2005. Data modelling and the formation of a grid. In *The node for secondary processing*, ed. John Kallas, 32-45. Athens: National Centre for Social Research.

Mejer, Lene. 2003. Harmonization of socio-economic variables in EU statistics. In *Advances in cross-national comparison: a European working group for demographic and socio–economic variables*, ed. Jurgen H.P. Hoffmeyer–Zlotnik and Christof Wolf, 67-85. New York: Kluwer Academic / Plenum Publishers.

Robson, Colin. 2007. *Real world research: a Resource for social scientists and practitioner-researchers*. Oxford: Blackwell.

Sarris, Willem E. and Irmtraud N. Gallhofer. 2007. Can questions travel successfully? In *Measuring attitudes cross–nationally: Lessons from the European Social*

Survey, ed. Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 53-74. Los Angeles: Sage.

Sundgren, Bo. 2003. Developing and implementing statistical metadata systems. http://www.epros.ed.ac.uk/metanet/deliverables/D6/IST-1999-29093-D6.doc

Zürn, Michael. 1998. Politik jenseits des Nationakstaats. Frankfurt am Main: Suhrkamp.

Notes

1 Kallas, Ioannis ,University of the Aegean, Department of Sociology ,Academic Field: Methods & Information Techniques of the Social Sciences.Address: Tertseti & Mikras Asias str., 81 100 Mytilene, Lesvos, Greece./ Tel. (+30) 22510 36559. email: i.kallas@soc.aegean. Linardis, Apostolos . National Centre for Social Research . Address: 14-18 Messoghion Av., GR-115 27, P.O.B 142 32, Athens,Greece.Tel. (+30) 210 7491656.email: alinardis@ekke.gr

2 http://www.europeansocialsurvey.org/

3 http://www.issp.org/

4 The documentation of all universe-specific entities is completed by all participating research groups, in the languages that have been decided by each group but also in the common agreed language.