Building an Infrastructure for Content Harmonisation and Conversion

Work Package 9 of the 2008-2009 CESSDA PPP under the 7th Framework Programme.

(See http://www.nsd.uib.no/cessda/project/index.html for information on the CESSDA PPP.)

With the widespread availability of cross-national survey programmes covering most European countries, comparability and harmonisation issues have become common to many areas of social research. Even in the preferred, but unusual, situation of data having been fully harmonised *ex ante*, as is the case in the European Social Survey, comparability remains problematic when it comes to comparing data across *different* survey programmes. An impressive number of recent and continuing harmonisation efforts relating to aspects of social structure underline the high demand for harmonised data in the research community, including the relevant departments of Eurostat and national statistical offices. Relevant examples include, among many others, ESeC on classes, Treiman's and Ganzeboom's work on socio-economic status and mobility, EU-SILC research on income harmonisation strategies (CHINTEX), and the planned "EDACwowe" portal for the "RecWoWe" project under FP6 on issues of work and welfare organisation.

What this series of independent efforts is mostly lacking, however, is a perspective to preserve and maintain the results over a time span that is much longer than the respective projects' duration. From the secondary analyst's point of view, a large part of the distributed harmonisation investments is currently at risk of becoming *inaccessible*. Equally important is the problem that information from these individual research projects is not necessarily *visible* to the wider research community.

In response to these concerns, the key purpose of the present work package is to develop strategic options and future directions for preparing the CESSDA Research Infrastructure to become a central focus for the collection, preservation, distribution, and further enrichment of such information, including the generation of additional harmonised datasets and tools for their discovery. The procedures and software products to be created in the subsequent implementation phase will become extremely useful tools in the harmonisation of existing datasets. At the same time, they will facilitate the design of input-harmonised surveys from the start. The infrastructure will, once established, be open to all of the European Research Arena. As the example of the ESS has shown, freely accessible high quality harmonised data can be expected to trigger a wave of contributions to European comparative research, substantive and methodological. Extending the range of such harmonised data via the new infrastructure will in particular foster contributions from individual researchers without special funding for harmonisation efforts. This will also invite methodological research about alternative measurement concepts for similar theoretical constructs found in survey programmes other than the ESS, for example EVS/WVS, CSES, and ISSP.

While the preservation and redistribution of data generated by primary researchers are natural tasks for the CESSDA member organisations, the European comparative nature and the sheer scale of the task at hand dictate that this must be undertaken via a concerted effort between data management experts and research networks. With these concerns in mind, this work package sets out to work on two substantive key objectives.

Objective 1 – To strategically plan for meeting substantive harmonisation demands in the European SSH research community.

In addressing this wide-ranging objective, as a starting point, in order to understand more fully the demands of the European research community, there is a need to generate comprehensive information on the motives and rationales that lie behind different existing harmonisation efforts

of comparative data. There is a requirement, therefore, to consult experts on substantial and methodological aspects of classifications and standards. A workshop with established experts from within and outside CESSDA will serve to identify the demands for standards, classifications and conversions, and to identify where these demands can be matched with existing survey materials (questions and data sets, including, for example, household surveys from official statistics). Contributors to recent or current harmonisation projects will be invited with priority, not only because of their special experience, but also to assess strategic options of joining efforts with and obtaining input materials from these existing projects.

Within this objective a second major theme will be to develop strategies to collect and create classification and conversion information. This will require the specification of sources and procedures for collecting classifications and scales and other 'raw materials' for harmonisation. A particular challenge will be to devise organisational concepts of how contributors of raw materials can input these materials into the front-ends of the harmonisation infrastructure, and how users can discover and access the contents of such an infrastructure.

Several options for collecting content contributions to the harmonisation infrastructure should be explored, including, for example, a Wiki-like community based principle, a closed circle of project financed experts, or contributions from national statistical offices and long-standing EU-financed programmes. A central question will be whether, and if, what incentives can be found that invite voluntary contributions of material.

Objective 2 – To develop functional specifications of infrastructure elements.

Harmonisation work involves processing and organising large amounts of information on materials such as constructs, existing classifications, and the related survey questions. This information needs to be organised along several dimensions. It is anticipated that collecting and managing such information will require a specialised technical infrastructure, as will distributing the compiled classifications, *et al* to the wider scientific community. Drafting workable functional specifications for such an infrastructure constitutes a major challenge, and creating such specifications is the core of the present project. The structure presently envisioned is that information on constructs and classifications is managed within one database (the Constructs, Classifications, Conversions DataBase: CCCDB), while information on questions and their source surveys is held in another (the Question DataBase: QDB).

A natural base of a harmonisation information system is a database for collecting concepts, classifications and conversion/harmonisation information. Examples of such databases are rare at best. To some degree Eurostat's RAMON 'metadata server' which collects definitions of international standards mostly from the field of economics and official statistics, can be seen to illustrate one aspect of the proposed concept, namely that of a central warehouse of classifications. But all existing examples either do not specifically address broader social research needs, or do they not directly support harmonisation and conversion work. Thus, a functional specification for a CCCDB is needed that should allow for linking all the elements of the harmonisation process, from survey questions at the input side, through the actual coding of variables in datasets, to a completed comparative classification or scale at the output end. Subsidiary information useful in the harmonisation process can, for example, be provided through links from constructs and classifications to articles on their theoretical foundations or methodological aspects with relevance to comparability.

Different types of classifications and conversion demands constitute very heterogeneous data structures and data types for the CCCDB. Even if much of the initial content shall be derived from the field of socio-demographic information, where most current harmonisation research has its focus, the architecture must be flexible enough to accommodate other types of

sociological measurements, such as behaviour self-reports in time-use surveys, and, of course, attitude or value scales.

A need for conversions not only arises with respect to comparisons across countries, but also across time. Not only scientifically designed measurement or classification systems change over time (for example, ISCO-68 was replaced by ISCO-88, which in turn will be superseded by ISCO-08; attitude scales may be partially revised sometimes), but also apparently 'natural' categorisations like geographical location identifiers or the borders of administrative regions undergo changes along with political developments. The CCCDB must therefore also reflect a time dimension and a version dimension to support trend analyses and historical comparisons.

Alongside the planned CCCDB a closely related functional specification is required for a Question DataBase (QDB). The QDB should perform two basic functions. The first is to inform secondary analyses in comparative research in an unprecedented way. This will make the QDB a highly relevant tool for any kind of *ex ante*- and *ex post*-harmonisation work. At present, if a researcher wants to access national versions of a question asked in a comparative survey, she or he has to browse through the field questionnaires of all relevant countries one by one. An essential capability of the QDB should be to present different versions (national or time-specific) of a question in a display that allows for immediate comparison. The quality of secondary analyses will be increased by facilitating the interpretation of national operationalisations; methodological reasons for differences in national results can be identified more easily. To allow such systematic comparisons, the design of the QDB must provide for searching and accessing individual questions through the concepts and classifications managed in the CCCDB.

The second basic function of the QDB should be to support the development of new questionnaires by giving researchers access to a selection of viable operationalisations of a construct. The same possibilities for easy comparison as described above will save time and effort in creating new surveys that can thus more easily continue time series started with specific questions in older surveys. Being able to check on a number of different translations will also allow more informed decisions about how to transfer questions into a new language.

Both databases' designs will have various aspects in common. First and foremost stands the requirement to have an interface to each other, allowing questions to be searched by reference to constructs, scales or classifications and *vice versa*. Secondly, the architecture for both must allow input of content from a large circle of contributors. In the long term, information from both databases is to be linked with general data distribution platforms (such as future versions of NESSTAR©). To meet this requirement, the new infrastructure elements must interoperate with standardised metadata documentation systems.

Participation in Work Package 9 (contact persons at the national partner archives)

The project team of Work Package 9 is comprised of staff from seven CESSDA members. Contact persons at each institute:

Markus Quandt (GESIS-ZA, Work Package Leader), markus.quandt@gesis.org
Nanna Floor Clausen (DDA), nc@dda.dk
Laurent Lesnard (CDSP), laurent.lesnard@sciences-po.fr
Jindrich Krejci (SDA), jindrich.krejci@soc.cas.cz
Tolis Linardis (EKKE), alinardis@ekke.gr
Hilde Orten (NSD), hilde.orten@nsd.uib.no
Marion Wittenberg (DANS), marion.wittenberg@dans.knaw.nl