The Contextual Database of the Generations & Gender Programme: Harmonized Contextual Data for the Analysis of Demographic Decision-Making

Arianna Caporali[1], Sebastian Klüsener[2], Gerda Neyer[3], Sandra Krapf[2], Olga Grigorieva[2]

1. Introduction

Demographic changes, such as continuing population ageing and decreasing fertility rates, are posing policy challenges to national governments in Europe and other developed countries. To meet these challenges, advances are required in the understanding of socio-demographic trends and of factors that influence these developments. Such undertakings need access to cross-country comparative individual data on demographic behaviour and to information on contextual political and socioeconomic conditions, in which this behaviour is embedded. However, for researchers it is often a tedious and time-consuming endeavour to compile cross-country comparative contextual data. Data has to be derived from different international and national databases and checked for reliability and comparability. Where cross-country comparable data for a specific indicator is not available, researchers need first to harmonize the available information before they can use it in their analyses.

The Contextual Database (CDB) of the Generations and Gender Programme aims to support researchers by providing harmonized cross-country comparable data on demographic, socio-economic, and policy contexts for up to 60 countries in Europe, Asia, North America and Oceania. The CDB is an integral part of the Generations and Gender Programme (GGP), which seeks to enhance the understanding of the development of fertility and family behaviour in Europe and beyond, and of the demographic, social, economic, and political factors that influence it.[4]. The CDB was developed to complement the individual-level data generated by the Generations and Gender Survey (GGS), a national panel survey conducted in intervals of approximately three years. Between 2009 and 2012, the GGP has received funding from the European Union (EU) 7th Framework Research Programme. Part of the funding has been used to develop and extend the CDB into a comparative database with harmonized economic, demographic, social, and political indicators. The contextual data provided by the CDB is now suited for cross-country comparative multi-level analyses. Beyond the GGS the CDB data can also be used as contextual information for analyses with data from other surveys (e.g. European Social Survey, SHARE). In addition, the database is of value on its own, as it constitutes an important data source for macro-level trend analyses. The work on the CDB is co-ordinated by the Max Planck Institute for Demographic Research (MPIDR) in Rostock (Germany). The database is freely accessible on the webpage of the GGP (http://www.ggpi.org/contextual-database.html).

In the first part of this paper we will present the conceptual considerations which have guided the set-up of the CDB. The second part provides background information on the guidelines which have formed the basis for the collection and harmonization of the data.

---

[1] Service des enquêtes et sondages, Institut national d'études démographiques (Ined, France).
[2] Max Planck Institute for Demographic Research (MPIDR, Germany).
[3] Stockholm University, Demography Unit (SUDA, Sweden).
[4] The GGP was launched by the Population Unit (PU) of the United Nation's Economic Commission of Europe (UNECE) in 2000 (United Nations 2000; Vikat *et al.* 2007). The EU project and the GGP programme are steered by a consortium of 12 institutions. Since 2009, the Netherlands Interdisciplinary Demographic Institute (NIDI) has been in charge of the co-ordination of the project.

2. Conceptual framework and content of the Contextual Database

A four-way approach guided the development of the conceptual framework and content of the CDB[5]. First, the content of the GGS questionnaire served as a starting point to determine relevant contextual domains (Spielauer 2004). These include economic, social, cultural, religious, educational, policy, and other indicators which may affect individual choices. Second, Neyer (2003) suggested to categorize these indicators along four theoretically and empirically meaningful dimensions: (1) equality; (2) agency; (3) social rights, and (4) risks and security. Third, to enable multilevel comparative studies in combination with GGS micro level data, the design of the CDB had to correspond to the retrospective, prospective and geographical design of the survey (Racioppi and Rivellini 2002). Fourth, inventories of existing international comparative databases (Bisogno 2002; Neyer 2003; Spielauer 2004) informed about data availability and past experiences in conceptual framework development and data collection.

The combination of these four approaches led to the identification of more than 200 variables structured around sixteen key topics: (1) general demographic indicators, (2) general economic & social indicators, (3) labour market and employment, (4) pension system, (5) parental leave institutions, (6) childcare policies and institutions, (7) military and alternative civilian service system, (8) unemployment, (9) tax/benefit system, (10) housing market and policies, (11) legal regulations of personal relations & family responsibilities, (12) education system, (13) health, (14) elderly care, (15) political system, (16) culture & values. The database consists of three main data types: 1) *National-level time-series*, i.e., time-series data that meet the historical depth of the retrospective component of the GGS; 2) *National-level policy histories*, i.e., policy history data that record key policy changes over time, and that can consist of both numerical information (e.g., replacement rates, durations) and short text descriptions; 3) *Regional-level variables*, i.e., cross-sectional information and short time series that record recent trends at sub-national levels (regions, provinces).

3. Data Collection

Up to 2008 data collection was conducted in a decentralized way by national teams of statistical offices, research institutes or research departments within statistical offices, which were involved in the GGP. The decentralized data collection caused challenges to the aim to compile cross-county comparative data. Another challenge was the limited functionality of the database in place.

The financial support received from the EU 7th Framework Research Programme allowed us to address and overcome these challenges. The CDB coordination group[6] started off with a comprehensive variable-by-variable comparison of already collected data for: cross-country comparability, completeness of the time-series, errors, deviation from the required definitions, and completeness of data sources. Additionally each variable was compared with variables in international databases of supranational organizations (e.g. European Union, World Bank, UNESCO, OECD, WHO) or research consortia (e.g., Human Fertility Database, and Human Mortality Database). The aim of this procedure was to figure out to what degree it was possible to complement collected data with data from international sources. A schematic overview of these comparisons was worked out. Two main sets of improvements derived from this work. On the one hand the guidelines for national data collectors were

[5] The CDB was developed on the basis theoretical and methodological papers prepared by the GGP-CDB Working Group. The group was headed by Patrick Festy from Ined (Festy 2004). For a complete list of individuals and institutions involved in the development of the database see the section "CDB About and Team" of the database webpage (http://www.ggp-i.org/contextual-database.html).

[6] The coordination team was composed of Arianna Caporali (data harmonization and documentation, review of national data collections), Sebastian Klüsener (relations with the national data collectors, conception of the new web environment, advisor in data harmonization and documentation), Gerda Neyer (senior researcher coordinator), Sandra Krapf (coordination of student helps), Olga Grigorieva (legal aspects linked to data dissemination), Fred Heiden (computer programmer).

improved so that they focus on collecting CDB national data, which is not available in existing international databases. On the other hand, the data harmonization and data preparation process by the CDB coordination team was modified. This was necessary to assure the best comparability over as many countries as possible and over as long time-periods as possible. To this end we also decided to enhance the transparency of the data sources, the collection, and harmonization process, by providing detailed metadata-documentation of each single data entry in the database. Finally, a new database environment was created to offer enhanced user friendliness in accessing and extracting data from the database. The new procedure of data collection now consists of five phases.

1. First, for a given indicator, we pull together all the available data and metadata in the same spreadsheet file.
2. Second, we cross-check all the different sources and select the best combinations of them. The choice of the data sources is determined by the following set of pre-established criteria: compliance with GGP-CDB guidelines and international standards, comparability across countries, completeness, spatial and temporal availability of the respective indicator, availability of well documented metadata information and variable definitions. Thus, CDB time-series may be the result of combinations of different data sources. If so, the following applies: 1) we prefer *national sources* provided by CDB national data collectors, whenever they are available and in compliance with our pre-established set of criteria (e.g., demographic indicators); 2) we prefer *databases of international organizations and/ or research consortiums*, in case of indicators that are already harmonized and checked for comparability across countries by these institutions (e.g., macro-economic indicators and labour market variables).
3. Third, we organize the metadata information. For each single data entry, we document not only the data source, but we also indicate whether any calculations or estimations were carried out by data producers, and whether there is any deviation from the variable definition and/or a break in the series.
4. Fourth, the collected data and metadata is uploaded in the new database web environment.
5. Finally, the harmonized time-series may be revised with the submission of new data collections by national teams.

4. Web interface and database functionality

The new database environment is set up as a dynamic system, based on a relational database (MS SQL Server). The web interface is programmed in PERL using additional technologies (JavaScript, Ajax and Flash). In contrast to the old static system it offers a dynamic choice of indicator values across countries, regions, time as well as other selection features, if available (e.g., age, sex). In addition, the user can choose the dimensions of the output (e.g. organize data columns by regions or by time, etc.) and several plot options (e.g., bar plot, line plot, pie plot). These plots are interactive, allowing the user to zoom in specific time periods or to in- or exclude countries and/or regions. Other new features are metadata documentation by single data entry and a geocoding option (see Conclusion for details).

5. Data availability as of November 2012

In accessing the CDB-webpage (http://www.ggpi.org/contextual-database.html), the user can choose between two options: The Contextual Database (CDB) and the Contextual Data Collection (CDC). With few exceptions, the CDB contains only harmonized contextual variables. As of November 2012, the database provides 93 indicators, covering up to 60 countries, in Europe, Asia, North America and Oceania. The time-frame reaches as far back as possible (for some indicators back to the 19th century) and ends with the most recent data available. The CDC contains the complete national datasets with more than 200 indicators, which is derived from the national experts in the participating GGP-

countries. The data in these national datasets are not always comparative across countries and they may not have been updated to the latest available date. But they are very rich in terms of national sources used and may be very useful for regional comparisons within countries. In total, the CDC contains eleven datasets available for download: Austria, Bulgaria, Canada, Germany, France, Georgia, Hungary, Lithuania, Norway, Romania, and Russia.

6. Conclusion

The GGP-Contextual Database comprises a unique combination of features which sets the database apart from the majority of other databases:

1. The CDB provides easy access to harmonized cross-country comparative time-series of demographic, socioeconomic and policy indicators. Often, this data is not only made available for the national level, but also for the sub-national regional level. This data can be used for macro-analyses and as contextual information for multi-level analyses of demographic and socioeconomic decision making by linking it to data from the Generations and Gender Survey or data from other surveys (European Social Survey, SHARE).
2. The CDB includes descriptions of key policy reforms in a large spectrum of policy domains.
3. The CDB provides metadata not only indicator-wise, but for each single data-entry.
4. The CDB makes harmonized time-series available in a dynamic user-friendly web environment which has innovative functionalities both in terms of metadata documentation and automatic geocoding of national as well as regional data. User can choose to include an ID-column in the output providing the geocode used in the GGS survey to identify the place of residence of an interviewed person. Through this code it is easily possible to match extracted CDB-data with the GGP-survey data. Also other regional coding schemes such as NUTS and ISO are supported, potentially allowing researchers to match the CDB-data also with data from other surveys. Data can be exported in different formats (e.g. CSV, XLS and XML).

The coexistence of all these features in the GGP-CDB makes it a unique support tool for researchers interested in the micro-macro linkages of social structures and processes. In addition, it offers high potentials to serve as a role model/platform for the development of future contextual databases for other surveys and/or world regions.

7. References

Bisogno, E. (2002). UNECE data of possible interest to GGP Contextual Database. In *GGP Contextual database group: Activities and conclusions* (23-26).

Festy, P. (2004). GGP Contextual database group: first discussions, first conclusions. Available from the author: festy@ined.fr

Neyer, G. (2003). Gender and Generations Dimensions in Welfare-State Policies, *MPIDR Working Paper* WP 2003-022. Rostock: Max Planck Institute for Demographic Research.

Racioppi, F. and G. Rivellini (2002). The Contextual Dimension in GGP: Some Methodological Issues about Data Collection and Sampling Procedures In *GGP Contextual database group: Activities and conclusions* (11-17).

Spielauer, M. (2004). The Contextual Database of the Generations and Gender Program: Overview, Conceptual Framework and the Link to the Generations and Gender Survey, *MPIDR Working Paper* WP 2004-014. Rostock: Max Planck Institute for Demographic Research.

United Nations (2000). *Generations and Gender Programme: Exploring Future Research and Data Collection Options*. New York/ Geneva: United Nations Publications.

Vikat *et al.* (2007). Generations and Gender Survey (GGS): Towards a better understanding of relationships and processes in the life course. *Demographic Research* 17 (14): 389–440.