

EIGE's Gender Statistics Database: Overview and Priorities

Background document for the experts' meeting

As part of EIGE's efforts in the collection and dissemination of gender statistics, the Institute's mid-term programme 2013-2015 tasks EIGE with developing and making publicly available a centralised **Database on gender statistics** (henceforth Database). The Database serves to coordinate, centralise and disseminate research data and statistics on gender equality in Europe. The Database supports better informed policy making at EU and Member State levels, and facilitates structured and user-friendly access to gender statistics information for users such as policymakers, civil servants, statisticians, research organisations, social partners, civil society organisations, media and, ultimately everyone living in the EU.

In this document, we provide a general overview of the database, set out future plans, and pose a series of questions for discussion.

Overview of the Database

Technical structure and organising principles

EIGE's database on gender statistics is a collection of statistical data and associated metadata pertaining specifically to the area of gender statistics. The logical structure of the Database is based on the **SDMX** (Statistical Data and Metadata Exchange) standard, which is an international standard for the organisation, production, and exchange of statistical information (data and metadata) among various data providers and users.¹

Datasets

The basic (lowest-level) organising elements ("building blocks") of the database are **datasets (DS)**. While in general there the term "dataset" tends to be used in a number of different meanings throughout statistics, we follow SDMX's and Eurostat's convention in defining a dataset as a set of observations that all meet the following two conditions: (1) they measure the same underlying concept (such as "employment", "employment rate", "level of education", "life expectancy", "satisfaction with life", etc.), AND (2) they are defined in terms of the same criteria (to be defined formally below, but loosely meaning the same breakdown variables, such as "sex", "age", "educational achievement", etc.). This second condition implies that "Employment by sex and level of education" and "Employment by age" are two separate datasets, unless observations also exist for intersections of age and level of education (e.g., separate observations for young college graduates and old college graduates). Note that even with this fairly narrow definition, the division of data into datasets is still

¹ https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Main_Page

somewhat arbitrary. For example, this definition allows to either store “Employment in 1000s of people” and “Employment rate (%)” in two separate datasets (each having “unit of measurement” as a dataset-level attribute) or in a single dataset, “Employment and employment rates”, which would then have an additional two-category criterion, which could be titled “Unit of measurement” (with “1000s” and “%” as categories) or “Indicator” (with “Employment in 1000s” and “Employment rate in %” as categories). In general, we adhere to the following rule: Whenever the immediate data source has provided datasets in SDMX format, we respect the source’s division of data into datasets (for example, Eurostat provides “Employment in 1000s” and “Employment rate” as separate datasets when the criteria are “time”, “country”, “sex”, “age”, and “nationality”, but it provides the same two measures as categories within an “Indicator” criterion when the only other criteria are “time”, “country”, and “sex”). In all other cases we group data at what we deem to be the most natural level for that particular collection of data.

Datasets are internally structured as follows: Each observation is a number linked to a set of qualifying **criteria**² (which both identify and describe the observation) and **attributes** (which only describe the observation). Each criterion has a finite number of possible values (**categories**). Together, the criteria form a multidimensional coordinate system, also known as a cube, where each point corresponds to exactly one category of each criterion. Each observation (a real number) is associated with a point in the cube. In addition, the attributes of this observation provide supplementary information that help interpret the number. These ideas are best described by example. Suppose our dataset represents a table containing the average annual employment rates for women and men in the EU-28 countries for each year between 2000 and 2013. Then, the data have three criteria: “Sex” (with two categories, “Men” and “Women”), “Country” (with 28 categories, each country being one category), and “Year” (with 14 categories, each year being one category). The resulting three-dimensional cube will therefore have $2 \times 28 \times 14 = 784$ points (observations), one for each possible combination of categories (such as (“Women”, “United Kingdom”, 2000)). Each observation will be the employment rate of the group defined by the corresponding categories (in our example, UK women in 2000). In addition to the criteria, each observation will have a series of attributes, some of which will be at the dataset level (the subject/concept measured (employment rate in our example) and unit of measurement (% in our example)), while others will be at the observation level (such as **flags** indicating whether the observation has been estimated or whether it represents a break in a series). Dataset-level attributes are part of the metadata for that dataset, while observation-level attributes are part of the data.

Currently, all observations in the database are at the country-and-year level, although we are working on extending the database to sub-national geographical units (regions) and sub-annual (monthly, quarterly, and biannual) frequencies. It follows that all datasets contain the criterion “Country” (to be renamed “Geographical area” after the extension to sub-national observations) and the criterion “Year” (to be renamed “Time period” after the extension to sub-annual frequencies). In addition, all datasets, except those that provide direct measures of the relative situation of women and men (such as values of the Gender Equality Index, gender pay gap measures, measures of gender gaps in other

² The SDMX standard refers to these as “dimensions”. In this document, we reserve the word “dimensions” for the dimensions of dataset views (to be defined below) and use the word “criteria” when referring to the general data structure.

variables (prepared for the Index), and indices of horizontal segregation in occupations and education) also provide sex-disaggregated data, i.e., contain the criterion “Sex”. The criteria “Country”, “Year”, and “Sex” (which we call the Defining Criteria) can therefore be viewed as collectively defining the basic unit of observation.

As stated in the previous paragraph, the database contains solely macrodata.³ However, most of these macrodata have been produced from microdata, either by the original or immediate source or by EIGE and its contractors. Depending on the way the data have been processed, we can distinguish three types of data: (1) data provided as macrodata by the source and used in the Database as-is (such as the employment rates), (2) data computed by EIGE or its contractors from other data provided as macrodata by the source (such as the gender pay gap on monthly wages, computed from levels of monthly wages that are obtained as country-level aggregates from Eurostat), (3) data computed by EIGE and its contractors from microdata (such as data on attitudes and opinions, computed from Eurobarometer microdata).

Dataset views

Datasets are displayed to users with the help of customizable tabular arrangements known as **dataset views (DSV)** or statistical tables. A dataset view is a two-dimensional layout presenting all or part of a dataset to the user. It is important to note that DSVs are the users’ only window into the data; there is no facility for users to examine datasets directly.

To define a DSV, one must specify one or more criteria as **row dimensions**, one or more criteria as **column dimensions**. The table is now formed by the intersection of (1) one row for each possible combination consisting of one category for each row dimension and (2) one column for each possible combination consisting of one category for each column dimension. The value displayed in any given cell is the observation pertaining to the categories defining the corresponding row and column.

Browsing tree

Dataset views are arranged in a **tree structure**, where they are grouped into several levels of **themes** (branches of the tree), with a small number of entry points constituting the highest level of **themes**. The current entry points are listed in Table 3. Any given dataset may have multiple associated DSVs linking this DS to a number of different themes (e.g., the employment rate datasets have a place both in the structure of the Thematic Areas and that of Policy Areas).

Table 1. The entry points of EIGE's database on gender statistics

1. Thematic areas

Under this entry point, data are organised according to general areas of interest. The structure is similar to the frameworks of Eurostat and national statistical institutes.

2. Policy areas

Under this entry point, data are organised according to the framework of European Union

³ **Macrodata** are statistical data observed at the level of countries or other geographical regions. This includes both data that are directly measured at the country level (such as GDP (Gross Domestic Product)) and aggregates (country-level statistics) of microdata (such as unemployment rate estimated by the EU Labour Force Survey (LFS) or public opinion as gauged by the Eurobarometer survey), while **microdata** are statistical data observed at the level of individuals, households, or firms (such as data from population surveys)

policy priorities. This entry point is specifically aimed at EIGE's main stakeholders, who are policy makers within EU institutions.

3. **EU strategies**

Closely related to "Policy areas", the "EU strategies" entry point organises data according to the priorities defined in a number of EU strategies, including "EU 2020" and the "EU strategy for equality between women and men 2010-2015".

4. **Gender Equality Index (GEI)**

This entry organises data according to the domains and subdomains of the Gender Equality Index.

5. **Beijing Platform for Action (BPfA)**

This entry point organises data according to the 12 areas of concern of the BPfA.

6. **Women and Men in Decision Making (WMIDM)**

This entry point presents data on the absolute and relative numbers of women and men in decision-making positions at the national and EU levels. It mirrors the DG Justice's database of the same name.

Metadata

Last but not least, **metadata** (or "data about data") are an integral part of the database, as the information contained therein makes it possible for users to understand, interpret, evaluate, and analyse statistical data. As explained earlier in this document, metadata are crucial for ensuring data quality along the Accessibility and Clarity dimension. There are two main types of metadata: (1) **structural metadata**, which provide a structured description of the way the statistical data and the reference metadata are organised, and (2) **reference metadata**, which provide additional descriptive information on the concepts used, the data collection and generation methods employed, and the quality of the data. Structural metadata essentially amounts to a formal definition and description of the data structure described above. The end user does not necessarily have to be aware of most structural metadata. Reference metadata consists of an extensive, mostly free-form description of the data, which allows the user to understand and evaluate various facets of the data, including (but not necessary limited to) the following:

- what the data purport to measure;
- how these measurements have been made;
- how the measurements should be interpreted;
- who is responsible for collecting and disseminating the data;
- how often the data are updated and disseminated;
- where the updated data and additional information can be found;
- how high the quality of the data is (within the framework described earlier).

In EIGE's database, structural metadata (such as the names and codes of datasets and dimensions, and the code lists of criteria and attributes) are embedded in the data, and reference metadata are presented in the Database alongside the data. The reference metadata follows the ESMS structure. ESMS "aims at documenting methodologies, quality and the statistical production processes in general. It uses 21 high-level concepts, with a limited breakdown of sub-items, strictly derived from

the list of cross domain concepts in the SDMX Content Oriented Guidelines (2009)".⁴ Below, we list the 21 top-level ESMS concepts:

Table 2. The top-level structure of the ESMS

1. Contact	2. Metadata update	3. Statistical presentation
4. Unit of measure	5. Reference period	6. Institutional mandate
7. Confidentiality	8. Release policy	9. Frequency of dissemination
10. Dissemination format	11. Accessibility of documentation	12. Quality management
13. Relevance	14. Accuracy and reliability	15. Timeliness and punctuality
16. Comparability	17. Coherence	18. Cost and burden
19. Data revision	20. Statistical processing	21. Comment

Technical implementation

EIGE's Database on Gender Statistics is a NoSQL database. A NoSQL database consists of a set of collections, each of which holds a set of documents. A document is JSON-style data structure composed of field-and-value pairs. Documents have dynamic schema, which means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data. The database stores documents on disk in the BSON serialization format. BSON is a binary representation of JSON documents, though it contains more data types than JSON. For the BSON spec, see bsonspec.org.

Current data content and data gaps

Data sources

The core of the database is based on data from reliable and comparable international sources. Where such data are scarce, they are supplemented by data from national sources (as is the case with administrative data on gender-based violence).

The following sources are currently included in the Database:

1. Eurostat's online database

All gender statistics from this database have already been included in EIGE's Database. This

⁴ See http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata/metadata_structure

includes all sex-disaggregated data and all data on gender equality issues (such as the gender pay gap).

2. EC DG Justice's Database on Women and Men in Decision Making

EIGE's Database currently mirrors the WMIDM database. EIGE is in the process of fully taking over the maintenance, updating, and development of that database, as discussed in a separate session.

3. The EU FRA study on violence against women

FRA's data currently form the core of EIGE's Database's section on gender-based violence

4. National-level administrative data on gender-based violence

Due to the lack of reliable internationally comparable data, EIGE has compiled statistics on the incidence of gender-based violence in EU Member States based on administrative sources in each State.

5. EIGE's own studies and computations

These are to be found under the Gender Equality Index entry point and under the entry point of the Beijing Platform for Action. They include the various components of the Index, as well as several indicators of the Beijing platform, for which no data are available from outside sources

The following sources are currently being processed for inclusion in the Database and should be available online by the end of September 2016:

1. General and Special Eurobarometer surveys

Active work is currently underway to process these important sources of data on attitudes and opinions. All General Eurobarometer surveys since 2000 and nearly 40 separate Special Eurobarometer modules are being processed. EIGE's Database will show country-level statistics (macrodata) computed at EIGE from original survey microdata.

2. Eurofound's studies: The European Working Conditions Study (EWCS) and The European Quality of Life Study (EQLS)

Five waves of the EWCS (1999-2010) and three waves of the EQLS (2003-2012) are being processed. EIGE's Database will show country-level statistics (macrodata) computed at EIGE from original survey microdata.

3. The EU LGBT survey

In 2012, EU FRA ran an online survey of LGBT persons' experiences of discrimination, violence and harassment. We are including macrodata from the published survey report.

4. The European Social Survey (ESS)

To supplement Eurobarometer's data on attitudes opinions, we are also processing data from the ESS (seven waves, from 2002 to 2014). EIGE's Database will show country-level statistics (macrodata) computed at EIGE from original survey microdata.

5. The OECD PISA survey

The Programme for International Student Assessment (PISA) is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of

15-year-old students. We are including data from the published survey reports (five waves, 2000-2012).

6. She Figures

Since 2003, the She Figures have monitored new developments related to careers, decision-making and, most recently, how the gender dimension is considered in research and innovation content. We are including data from the published survey reports (five waves, 2003-2015).

Data completeness and data gaps

As long as new and better data continues to be produced, the Database will never be fully complete. Instead, it is meant to be a project continuously under development and expansion. The database structure is also specifically designed not only to show all available data, but also to highlight the areas where major data gaps exist.

Currently, the area with the most severe gender gaps is the area of gender-based violence. The only internationally comparable data come from the joint Eurostat-UNODC crime statistics data and from the EU FRA survey on violence against women (which suffers from serious limitations due to the effects of differences in social norms on perceptions of violence). The available national-level data from administrative sources is both incomplete and incomparable (due to international differences in definitions and methodologies).

The overall state of the data (by entry point of the browsing tree) is given in the table below.

Table 3. Data completeness

ENTRY POINT	COMPLETENESS
Thematic areas and Policy areas	Most themes have an adequate amount of available data, although there is room for improvement. Areas with the most severe gaps are gender-based violence and attitudes and opinions. The latter gap, however, will be largely filled when the current survey data processing effort (see above) is complete.
EU strategies	Currently only two strategies are covered, "EU 2020" and "EU strategy for equality between women and men 2010-2015". More strategies need to be considered in the future.
Gender Equality Index (GEI)	The Database includes the scores of the Index and its subdomains, as well data for individual constituent variables. We are planning to expand the Related Variables section.
Beijing Platform for Action (BPfA)	All indicators with currently available data are fully

included in the Database.

The following indicators have no data, either because no data are available, or because the indicators cannot be quantified: D1, D2, D4-D10, E1-E4, F4-F7, F15-F17, G4, and J3.

The European Union has not defined any indicators under area I.

Women and Men in Decision Making (WMIDM)

The Database fully mirrors the WMIDM database maintained by EC DG Justice. As EIGE takes over that database, we plan to expand it with additional indicators.

Update procedures and schedule

The update procedures and schedule depend on the source and/or type of the data:

- For **Eurostat's macrodata** (which constitute the bulk of EIGE's Database):
 - At least once per quarter, the data and metadata in all datasets that have been updated with new or revised data or metadata on Eurostat's side (without changes to the dataset structure) are updated using data automatically downloaded and processed from SDMX files available in Eurostat's Bulk Download facility.
 - In addition, also at least quarterly, we look for structural changes in Eurostat's database (added, deleted, renamed, or restructured datasets) using an automated procedure based on the table-of-contents files provided in Eurostat's Bulk Download facility. The identified changes are then be examined by our experts to determine what structural changes to make to our database (datasets to be added, renamed, or deleted).

As an example, during the latest update to EIGE's Database, we added 199 datasets to EIGE's Database, renamed 59 datasets, removed 12 obsolete datasets, and restructured the section of the database related to education (20 datasets) based on classification system changes by Eurostat.

- For data on women and men in decision making positions (**WMIDM**):
 - Unit now, we have been updating these data from Excel files downloaded from the Commission's website
 - After the transfer of the WMIDM database to EIGE is completed and EIGE takes charge of collecting the data, data will be uploaded to the Database in SDMX format as soon as they are collected and verified.
- For data under the **Gender Equality Index (GEI)** and **Beijing Platform Action (BPfA)** entry points:
 - Many of the indicators are computed directly from Eurostat, WMIDM, or other external sources that are also used in the Database in their own right. In these cases, the

indicators are automatically recomputed (by the Database engine) as soon as the new source data are loaded in the Database

- Those indicators that are based on ad-hoc data collection or computation by EIGE (most notably, the Gender Equality Index scores) are updated when EIGE provides us with the updated data.
- For publicly available **macrodata** from sources such as SheFigures and the OECD PISA:
 - We monitor the sources' websites to identify new data releases
 - Once a new release is identified, the newly available data tables will be analysed to identify those that are relevant for the Database (paying special attention to data series that are repeated from previous years and can be reported in the same data tables as the earlier data).
 - The identified data will be downloaded, structured as multidimensional cubes and written in xml files conforming to the Database's internal. See Appendix 1 for details.
- For publicly available **microdata** from sources such as Eurobarometer, Eurofound, ESS, and EVS:
 - We monitor the sources' websites to identify new data releases
 - Once a new release is identified, the questions will be analysed to identify those that are relevant for the Database (paying special attention to questions that are repeated from previous years and can be reported in the same data tables as the earlier data).
 - For the relevant questions, statistics at the national level (macrodata) will be computed, structured as multidimensional cubes, and written in xml files conforming to the Database's internal format using the procedure described in Appendix 1.
- For **limited-availability** microdata (such as EU-SILC and EU-LFS):
 - We will work closely with EIGE and the original providers (most notably, Eurostat), to work out an algorithm and schedule for obtaining and processing the data in a manner that satisfies the source's data release and privacy policies
 - Once the microdata are obtained, we will follow the same procedure as for publicly available microdata (described above)

Keyword tagging system

EIGE's gender statistics database is designed to be navigated either by following the browsing tree or by using keyword (tag) search. When using the latter means of navigation, users search for datasets by typing a series of search terms (keywords, tags) into a text box.

Each dataset and each dataset view is tagged with a series of keywords. This is accomplished as follows. First, a set of valid keywords is defined for the whole database (this is necessary to avoid the automatic generation of meaningless keywords, such as articles and prepositions). Next, keywords are automatically linked to datasets and dataset views based on their names and descriptions. Finally, the generated keyword lists are manually edited to arrive at final keyword lists for all datasets and dataset views.

We are currently considering gradually moving to a significantly more advanced keyword system. There is going to be a dedicated working group for this purpose at the September 15 expert meeting. Please also see the relevant background document.

User interface

EIGE's gender statistics database is designed to be navigated either by following the browsing tree or by using keyword (tag) search. Whichever method is used to locate a dataset view, once it is selected, it is displayed in the browser window both as a visualization (graph) and as a data table.

A problem with the current interface is that it generally displays only a small portion of the dataset view at a time. This problem arises from the fact that the interface is visualisation-centred, not data-table-centred. The table can only display what is shown in the graph. This is different from the originally intended format, where by default the tabular layout predefined in the dataset view is displayed, and the user can then further customize the table to obtain any desired tabular representation of the data. Visualisations can then be requested as particular sections of the displayed table.

Transforming the interface to conform to this format is one of the activities we are planning for the future (see activities going forward below).

Activities going forward

Maintenance and updating process

Existing data series and metadata will continue to be updated following the procedures and schedule outlined earlier in this document (see section "Overview of the Database", subsection "Update procedures and schedule").

In addition, the database software will continue to do be maintained and updated. In particular, this includes monitoring of the database management system performance and taking the necessary measures to improve the performance, as well as updating the database management system with modules, updates and hotfixes which will be necessary for the smooth operation of the database, enabling and maintaining the connections with data sources.

We can envision a series of further technological improvements that could improve the user experience and ease the implementation of further content improvements:

- (a) improving performance to enhance the speed of the database (to be achieved through upgrading to the newest version of MongoDB, taking advantage of the performance-enhancing new features of this upgrade, analysing and rectifying issues with the current deployment configuration, and adding new indexes for more efficient search);
- (b) automating database backup and maintenance tasks;
- (c) improving data export functionality by adding automatically generated code to enable users to more easily import the extracted data into Stata and SPSS;
- (d) unifying the naming of criteria and categories that are currently named differently in different sources, which will improve search capabilities and will ultimately enable DS merge for

analyses spanning multiple datasets (such as creating scatter plots with data from different datasets on the two axes).

Development and possible expansion of the database

Data and metadata

Consideration of additional data sources and data series

We are constantly looking to expand the coverage of the Database, including as many sources as possible, subject to the quality requirements set out below.

The task of adding new data consists of two steps:

This task consists of two steps:

STEP 1: Identifying new potential sources of statistical information

This will be done by means of desk and web research, consultations with researchers in our experts' networks, consultations with statistical data providers, and analysis of input from expert meetings, user surveys, and online discussions.

STEP 2: Assessment and feasibility analysis of the identified new sources

Data sources and particular data products considered for inclusion in EIGE's gender statistics database are evaluated based on the **quality-assessment** framework delineated in Appendix 2.

In principle, most quality dimensions can be evaluated both qualitatively and quantitatively. However, given the resources available to us and the large body of data to be evaluated, we are currently relying solely on a qualitative assessment of the documentation provided to us by the original data provider.

This evaluation process is straightforward when the source has provided adequate ESMS metadata, in which case careful examination of the metadata is generally sufficient for an overall quality assessment. Otherwise, all available documentation will be scrutinized to create a quality assessment following the ESS framework. In the case of microdata in particular, questionnaires and sampling schemes will be examined to identify gender bias and limitations to accuracy due to sample under-coverage or insufficient sample size.

When a given statistic is available from multiple sources, the highest quality source will be selected. All else being equal, data provided by Eurostat will be given priority over alternative sources, due to the perceived most meticulous quality management procedures, specifically with respect to international comparability.

When a statistic is available from only a single source (such as most attitude measures from Eurobarometer and survey-based estimates of the prevalence of various forms of gender based violence from the FRA GBV survey), that source will almost always be included, as long as it meets a **minimum quality standard**: namely, **disaggregation by sex is available, and sufficient metadata are provided to identify the source and understand how and what is being measured**.

Given that these minimum standards are met, the data will be included in the Database, but, if the quality assessment shows serious inadequacies, the presentation of these data in EIGE's database will

be changed to reflect these inadequacies. This will be done on several levels: first, any such inadequacies will be made explicit in the appropriate section of the structured metadata provided in the Database; second, when the inadequacies are severe, they will be noted in the description displayed on top of the data table; third, when there are particularly severe problems with comparability across countries (as is the case with administrative data on gender based violence), the data tables for different countries will be displayed separately to discourage the users from direct comparisons when such comparisons are not warranted by the data.

In summary, then, the source selection algorithm for a given data item/ statistic is as follows:

1. Identify all sources for the given data item
2. Keep only sources that meet two minimum requirements:
 - a. sufficient metadata are available to identify the source and to determine how and what is being measured;
 - b. sex-disaggregated data are available (note: this requirement does not apply to gender-based violence and concepts that directly measure a gender equality concept, such as the gender pay gap).
3. Include in the Database the highest-quality source among those remaining
4. When the metadata reveals serious quality problems, change the presentation of the data to reflect these problems:
 - a. Do not display non-comparable data side-by-side;
 - b. Reflect problems in the metadata displayed along with the data.

Data quality evaluation from a gender perspective

The evaluation of data quality, **with particular focus on the gender perspective**, is both an important step in evaluating new data sources and an important activity to carry out to evaluate the existing body of data.

Quality is a complex, multidimensional concept that measures the fitness of data for their purpose. The identification of data quality with fitness of purpose is rooted in the belief that quality should not be assessed on entirely absolute grounds, but should rather be evaluated with respect to the intended users of the data and the uses to which the data are expected to be put. This is particularly relevant in the context of EIGE's database, which focuses exclusively on gender statistics, which are intended for the very specific and clearly defined purpose of measuring and advancing gender equality.

The general framework we use to evaluate data quality is borrowed from the quality assessment and assurance frameworks of the European Statistical System (henceforth ESS). Based on the European Statistics Code of Practice, the ESS quality assurance framework evaluates quality along three blocks of dimensions: institutional environment (principles 1–6 of the Code of Practice), statistical processes (principles 710), and statistical output (principles 1115).

For a detailed discussion of our approach to data quality evaluation, please see **Appendix 2**.

The results of the quality analysis conducted will be presented in two forms: first, a **full report** on the results of the analysis and, second, a summary of the findings in the relevant sections of the **reference metadata files** for the affected data.

Specific focus on two areas: gender-based violence and women and men in decision-making positions (further discussed in separate sessions)

In separate workgroups in the September 15 expert meeting, we will focus on two areas of specific interest: gender-based violence and women and men in decision-making positions.

The first of these two areas, gender-based violence, is an area that is of utmost importance to European policy making, yet suffers from a severe shortage of internationally comparable data. Filling the existing gaps in this area should be seen as one of the key priorities both in the wider programme of gender statistics production and dissemination in Europe.

The second area, women and men in decision-making positions is of particular interest because EIGE is in the process of taking over the Database on Women and Men in Decision Making (WMIDM), which has so far been maintained by European Commission's DG Justice. EIGE will be responsible not only for the processing, storage, and dissemination of existing data, but also for the production of new content, including regular updates of existing data series, revision of current methodologies, and development of methodologies for the collection of data in additional domains to be included in the database in the future. The dissemination of data will be accomplished via the WMIDM entry point of EIGE's Gender Statistics Database.

For details on these two areas, please see the relevant dedicated background documents.

User interface and presentation of content

Customizable pivot tables

The current interface of the Database is (we hope) generally intuitive and visually appealing. However, its present implementation **fails to adequately reflect all the content defined in the Database**. Specifically, the information coded in Dataset Views (DSVs) is not fully utilised. As discussed earlier in this document, DSVs define which information from a given dataset should be displayed by default and how the data should be arranged in tabular form. The incomplete treatment of DSVs results in irrelevant information sometimes displayed in the tables and visualizations and some pertinent information not being shown by default. The interface also does not allow customization of data tables. As a result, **the current interface is usable for visualisations, but not ideal for the presentation of data in tabular form**.

The current limitations of the user interface arise from the fact that it is visualisation-centred, not data-table-centred. The table can only display what is shown in the graph, which, in turn, is generally limited to a single time series. This is different from the originally intended format, where by default a complex, multidimensional tabular layout (predefined in the dataset view) is displayed, and the user can then further customize the table to obtain any desired tabular representation of the data. Visualisations can then be requested as particular sections of the displayed table.

Transforming the interface to conform to this format is one of the activities we are planning for the future. The desired end result should work as follows:

1. Upon loading, display the **default data table** arrangement defined in the dataset view (DSV).
2. Make the **default visualisation** reflect a section of the default table as defined in the DSV. It might also be advisable to extend the DSV definition so that it also explicitly specifies a default visualisation.
3. Allow users to **customize the arrangement of each data table** in a manner similar to that employed in the Contractor's preliminary data viewer and in Eurostat's online database.

Simplified tabular layouts of complex tables and improved table descriptions

Once the interface is updated to fully reflect the information reflected in the dataset views (DSVs), further work is planned to improve the presentation of data in the DSVs, particularly by simplifying tabular layouts when necessary and by providing more extensive table descriptions.

Recall that each dataset has one or more DSVs associated with it. Each DSV is tied to particular point in the browsing tree and defined how the data from the associated dataset should be displayed when the user browses to it via this part of the tree. The DSV defines the default tabular layout of the data and optionally provides a table description that succinctly summarizes the most important definitions and explanations from the associated references metadata (such as the theoretical and operational definition of the concept measured).

Given the large number of tables/ DSVs it is not possible to provide the above-mentioned concise table descriptions for all tables. However, the goal should be to continually increase the number of tables for which such descriptions are provided. Similarly, it is not possible to customize the default views of all tables, yet we should strive to continually add more simplified table views that would highlight the pertinent information in an easier-to-view form. Through consultations with both internal and external users (see below), we will also solicit suggestions for additional simplified views and improved descriptions to be included in future iterations of the development and maintenance of the Database.

Improved keyword system

The current keyword system does not take full advantage of the rich relationships between keywords referring to various gender equality concepts, general economic or social concepts, and statistical terms. Using pure, unlinked keyword tagging also misses the opportunity to provide definitions and richer background information for the search terms.

Fortunately, EIGE already has an instrument in place that has the potential to serve to address these inadequacies. This instrument is EIGE's Gender Equality Glossary and Thesaurus, a specialised terminology tool focusing on the area of gender equality. The glossary contains over 400 terms in English with their definitions and sources. It allows users to find comprehensive and detailed definitions and explanations of specific terms and to compare and contrast similar or otherwise related terms. The terms are organised in four domains and 12 subdomains.

By using the Glossary and Thesaurus as a starting point and making a few additions, such as adding a separate domain for statistical and economic terms, defining new types of relationships between terms (such as "is a measure of" or "is an alternative measure to"), and adding links from the terms to dataset keywords, we hope to build a lexicographic statistical information system that will allow users

to explore various gender equality concepts in depth, including both conceptual understanding and statistical measures and evidence.

We are holding a separate working group on the keyword system. Please see the relevant dedicated background document for details.

Dissemination/Communication process

Gathering the opinions of users and stakeholders

To continue serving the interests of all users and other stakeholders, we continuously strive to gather feedback from various groups of users, including policy makers, researchers, journalists, activists, and EIGE's staff. Feedback from data providers, particularly from Eurostat and national statistics offices, is also to be considered. Feedback will be obtained using the following channels:

- (a) experts' meetings;
- (b) online discussions with EIGE's stakeholders;
- (c) online surveys of Database users;
- (d) online monitoring tool to record database usage;
- (e) internal tracking of change requests at EIGE.

The issues on which feedback will be solicited include:

- (a) areas where data are lacking;
- (b) organisation of the browsing tree;
- (c) search capabilities of the Database;
- (d) typical use cases and ways to make these easier via improvements in the Database;
- (e) organisation and presentation of the metadata;
- (f) ways that the short information displayed along data tables could be improved;
- (g) any additional issues that will be identified during the consultation process with EIGE.

Regular preparation and dissemination of short publications on the Database and its contents

To aid the dissemination of the data, we propose to regularly publish two types of short online notes:

- Statistical notes analysing the situation of women and men (including trends and new developments) in the areas covered by the Database;
- Summaries of recent updates, extensions, and improvements to the content and functionalities of the Database.

Statistical briefs

These briefs will cover all the key areas of the database. There should be several of such briefs published every year. The exact coverage, format and content of these analyses is yet to be determined. This year, we have already produced such analysis for migration and health.

The methods used for developing the statistical briefs might include (but will not necessarily be limited to):

- qualitative research building on existing literature (policy documents and reports from the EU institutions, national governments and relevant stakeholders; studies, publications, as appropriate);
- statistical analysis on data extracted from the Database;
- statistical analysis of new data not yet included in the database.

Summaries of new developments

These summaries will be published once every six months. They will document all changes made to the database since the previous update, including both structural and functionality improvements and content/ data updates and extensions.

If new updates show marked changes in data, these will be highlighted by means of graphs and short analyses. If completely new data series have been published, graphical snapshots illustrating the most notable patterns and/or trends in these data will be presented.

On the basis of these summaries, we also expect to regularly update the leaflet on the main features of EIGE's Gender Statistics Database.

Partnerships and cooperation

EIGE will continuously strive to establish new partnerships for further development and communication of the database. Potential partners include:

- Data providers, such as national statistical offices, Eurostat, EU agencies and other;
- Multipliers in civil society, academia and other fora;

Questions for discussion

1. What are your general views of the Database and the direction we are heading?
2. Which are the key areas that are most in need of improvement?
3. How important do you think it is to be able to display complex tabular layouts? How important is it to have user-customisable pivot tables?
4. Are you satisfied with the overall structure of the browsing tree? How could we improve it? Should we consider increasing the number of entry points?
5. What are your thoughts on the data update schedule? How frequently should we aim to update the Database?
6. Which are the most important data gaps?
7. Can you think of any additional data sources that we are not currently employing, but should consider for the future?
8. Do you think we should move beyond the national level and try adding some sub-national/ regional data?
9. How often should we aim to hold expert meetings and online discussions?
10. What topics should we prioritise for the statistical briefs?

