STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS **WORKING PAPER No. 6**

Work Session on Statistical Output for Dissemination to Information Media
(Voorburg, Netherlands, 25-27 September 2000)

ENGLISH ONLY

THE CENTRAL DATABASE TEMPO – A WAREHOUSE SOLUTION FOR MACRODATA DISSEMINATION

Paper submitted by National Commission for Statistics of Romania [1]

Summary

Supporting the modernization of Romanian economy by provision of accurate and timely statistical information, a "Time Series Database" project was developed in NCS with technical assistance from PHARE funds.In a classical-by-now view of a Statistical Information System with 3 layers (Production, Reference and Dissemination), this database named TEMPO belongs to the Statistical Reference Environment, containing established data in a standardized, uniform description, associated with metadata.

The main characteristics and functionalities of the TEMPO database are presented. The database contains statistical time series with different periodicities, held in multi-dimensional matrices. Matrix dimensions are related to statistical nomenclatures and classifications. Major categories of metadata are added to the matrices. The system has support for textual information in many languages.

System functionality covers both data management and data consultation.The "Consultation" application is presented, going through matrix search, matrix extraction and display, matrix calculation, matrix export, matrix save and retrieve.Some project implementation issues and further steps are considered.

I. INTRODUCTION

1. In order to support the statistical operations of data production, analysis and dissemination, accross subject-matter areas, a need has been formulated for the development of a consistent, domain-independent system for the management of time-series.

[1] Paper prepared by Ms. Liana Marina, Project Leader, National Commission for Statistics (NCS), Bucharest, Romania.

2. The main objective of the system was the improvement of access to reliable statistical information both for internal and public domain users according to their requirements, through modern information technologies. Placing the Time Series Datebase (named TEMPO) into the Statistical Reference Environment, as a "unique, shared data source" leads to:

- minimizing the number of connections between the data dissemination department and the production units inside NCS;
- simplifying the data flows;
- better administration and control of data, since centralized;
- enforcing standards.

3. Some of the main purposes of the new system can be enumerated:

- calculation of derived indicators;
- data extraction for statistical analysis;
- providing statistical data for outside users;
- enabling methodological revision.

4. The new time-series database TEMPO was developed in the National Comission for Statistics between 1997 and 1999, in the framework of a technical assistance project financed from PHARE funds.

1. Main Features

5. The TEMPO database consists of statistical time series (with yearly, half-yearly or monthly, quarterly data) and related objects (classifications / nomenclatures, measurement units, methodological notes, data sources, etc.).

6. Dealing with time-related data is a requirement that leads to a complex model, where each object has attached an interval of validity and many checks have to be performed in order to keep version control and to provide continuity of the series.

7. The time series database TEMPO is held centrally, at the NCS headquarters. Data files output from statistical production processes are used as database sources; data loading is done in a batch way, data entry (in an interactive manner) and data updates being not widely used; a common interface between the data sources and the database is implemented, in order to unify the different structures of the source data files.

2.1 Multidimensional Matrices

8. Statistical indicators, no matter the statistical domain they come from (i.e. industrial statistics, labour force, education, population, national accounts, agriculture, trade statistics, tourism, etc.) have a multi-dimensional nature, i.e. they can be represented as multi-dimensional matrices, known also as "hyper-cubes" or "cubes".

9. For example, a possible breakdown of the statistical indicator "School population" can have 4 dimensions:

- type of education (preschool, primary, gymnasium, secondary,...);
- district of Romania (Alba, Arad, Arges,…);
- gender (M, F);
- time (1990, 1991,…, 1999).

10. If each of the dimensions has a numer of 7, 42, 2, 10 positions, respectively, then the cross-classification of the 4 dimensions gives a matrix of (7 x 42 x 2 x 10) values (or cells). The greater number of dimensions, the more detail data in the matrix.
Matrices are described in terms of their dimensions in a generic way.

2.2 Macro-economic Series

11. The values stored in the database are data validated in the statistical production process and they are in some way aggregates, not micro-data; this characteristic will enable statistical data to be more accessible for the public.

12. Nevertheless, confidential data can be stored as well and marked as such, if required. Distinctions are made between definitive, revised and provisional values.

2.3 Nomenclatures and Classifications

13. As matrix dimensions are based on nomenclatures and classifications, the **TEMPO** database has a "Nomenclature Management System" component, dealing with all the nomenclatures related to (and shared by) the matrices stored; this component covers the nomenclatures and nomenclature items description and characteristics over time, through a generalized model, and should enable the consistency between statistical matrices, as much as possible.

14. Nomenclatures and classifications attached to matrices, as dimensions, are of two categories:

- nomenclatures of general use (standard);
- nomenclatures specific to a certain statistical domain.

15. Nomenclatures used in statistical production have also different degrees of harmonization with European and international standards.

2.4 Metadata

16. The database must also include necessary metadata for documenting the time-series and for supporting end-user access to the data; the main categories of metadata required are:

- subject-matter areas;
- indicators definitions;
- scope and methodology;
- breaks in matrix;
- footnotes for the matrix cells;

- contact person for additional information about indicators.

2.5 Data Security

17. A data security and access control policy was established and implemented, based on user and user groups permissions. The users groups are:

- the database administrator, responsible for the management of data and metadata;
- the NCS users (statisticians), having access in consultation to confidential or public data and metadata;
- the public, having access in consultation to public data and metadata.

2.6 Multi-lingual

18. The system has support for textual informations in many languages, both for database content and for the user interface.

2.7 Hardware and Software

19. The infrastructure supporting the system is a client/server network with UNIX (ICL NX) / Windows NT servers and Windows 95 workstations as clients.
The software tools used are: Oracle 7.3 database as server engine and Delphi 4.0 (database consultation), Oracle Developer/2000 (database management) as development environments for the client applications.

3. "Consultation" Module Functionality

20. The following main functions are developed in order to allow user access to the database content:

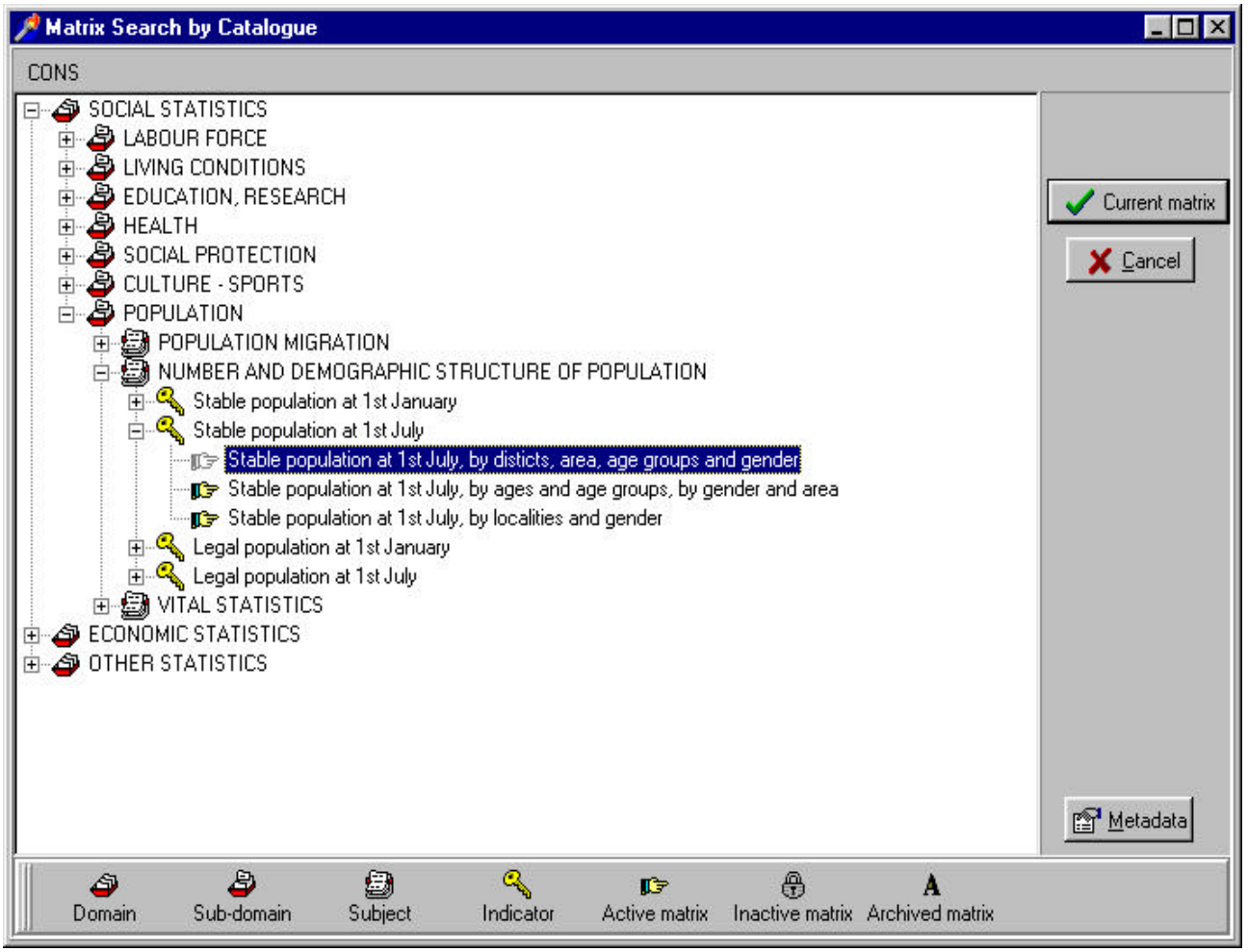
3.1 Matrix Search

21. A number of search criteria were implemented for locating a multi-dimensional matrix in the database, i.e.:

3.1.1 Search through the Catalogue

22. The Catalogue is a 5 level classification for retrieving a matrix, according to the statistical subject-matter area, namely the domain, the sub-domain, the subject, the statistical indicator, the matrix.

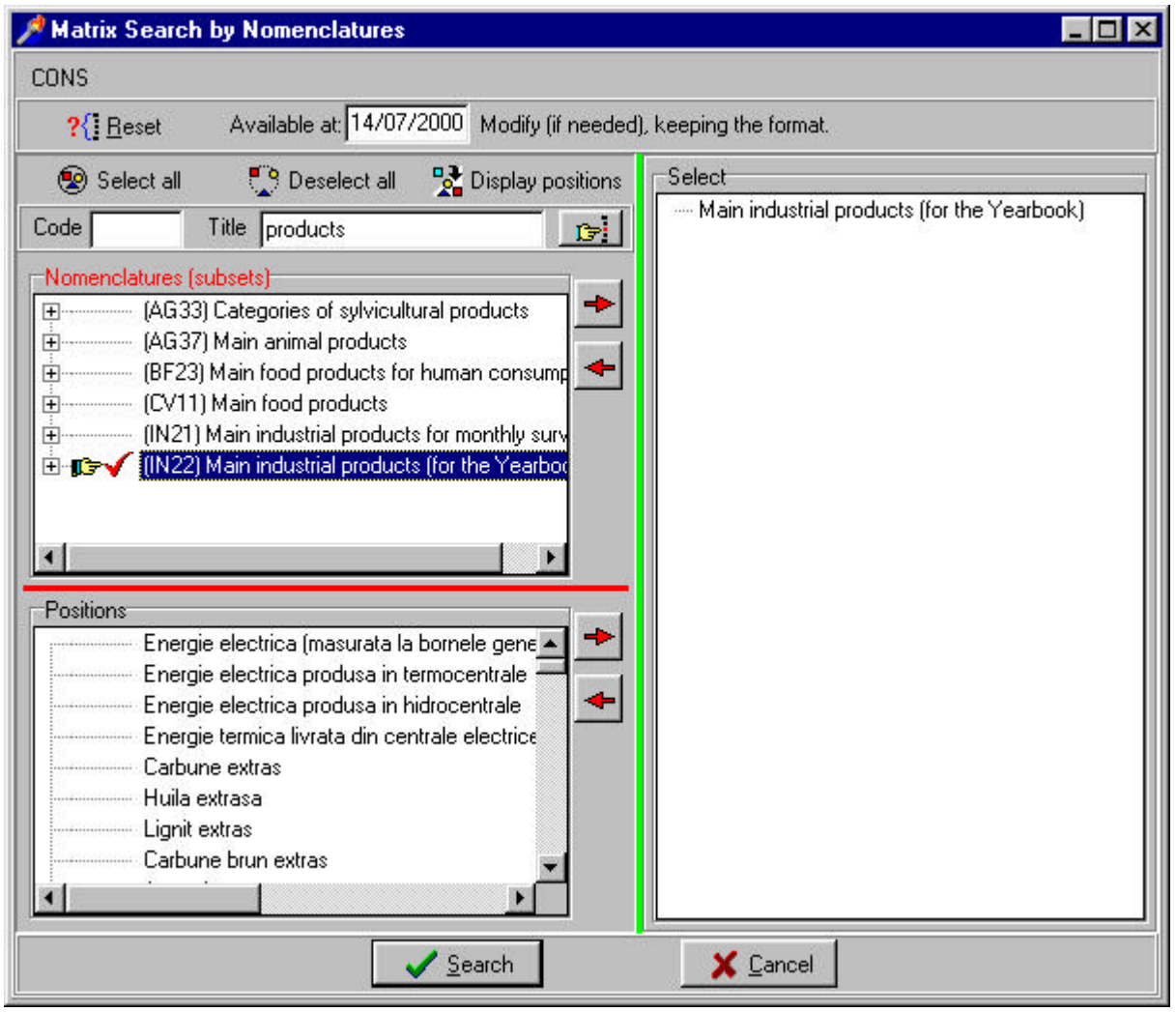
23. The Catalogue has currently 3 domains, 15 sub-domains, 48 subjects, 346 indicators and 537 matrices.



3.1.2 Search by Nomenclatures

24. Matrix searching is done by reference to one or many nomenclatures or nomenclature elements, such as level, subset, position. Over 250 nomenclatures with approximately 9200 positions are used for the description of matrix dimensions.

25. The list of matrices linked to either one of the selected items is shown. One matrix can be selected for further operations.



3.1.3 Search through Keywords

26. A string of characters is searched against the matrix name or the indicator name; an exact match, non case sensitive, is required. The list of matrices fulfilling the condition is made available.

3.1.4 Search by Periodicity

27. Matrices that have a certain periodicity of their values are retrieved.

3.1.5 Search by Data Source

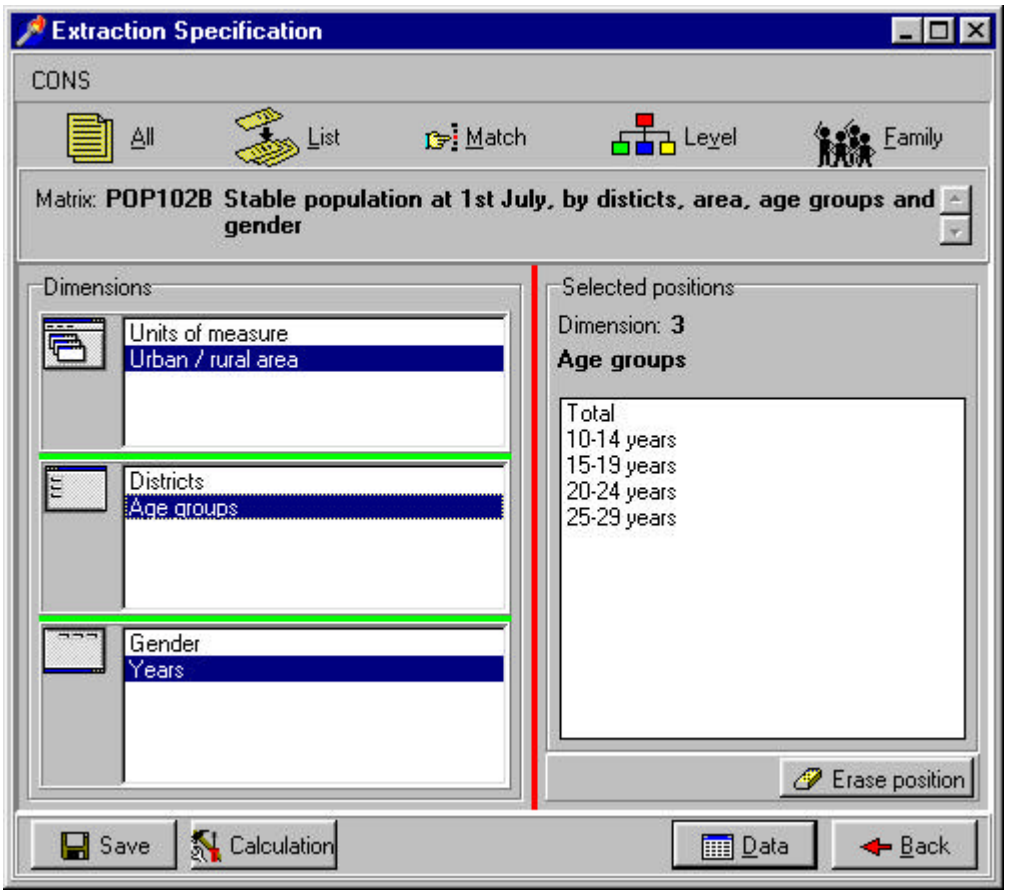
28. Searching is done starting from the list of statistical surveys or of external sources, such as Ministries, Administration organizations, etc. Matrices that contain data from the mentioned sources are retrieved. A number of 117 different sources for the data are identified.

3.1.6 Search by Indicator Code or Matrix Code

29. A pattern for the indicator code (6 character string) or the matrix code (7 character string) is required. If the indicator code or the matrix code in the database are known by the user, this is the quickest way of retrieving a matrix.

3.2 Matrix Extraction and Display

30. In order for a matrix to be displayed, an extraction of its cells needs to be done. By default the whole matrix is extracted, but in the case of long time-series, a matrix can count tens of thousands or hundreds of thousands of cells.



31. Matrix extraction is done according to an extraction specification, meaning a selection of positions in each of the matrix dimensions. Only data corresponding to the selected positions are extracted from the database and displayed.

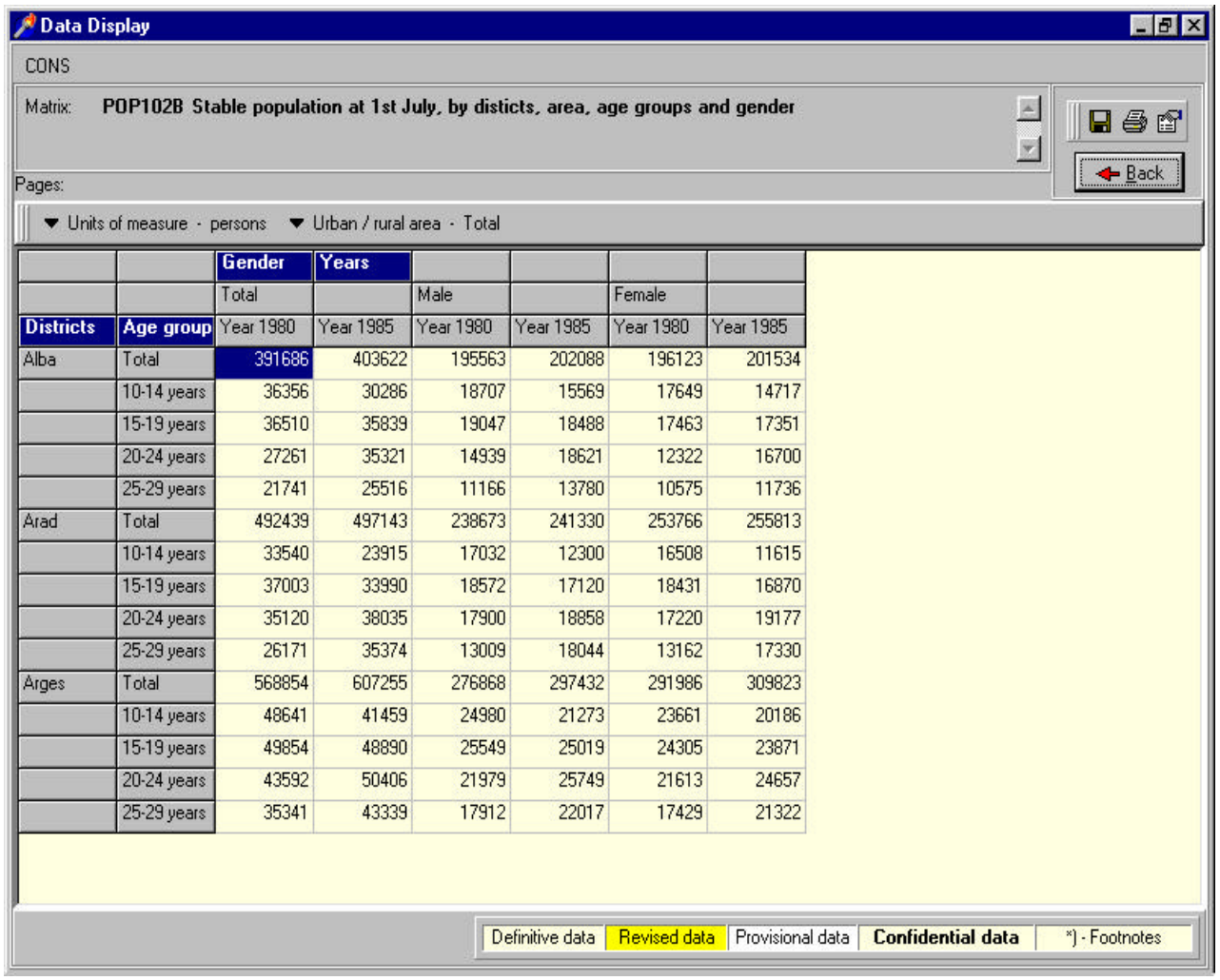
32. Each matrix has maximum 12 dimensions, two of them being mandatory dimensions: the “Time” and the “Measurement unit” (the matrix in the example above has 6 dimensions).
From a matrix display point of view, the matrix dimensions fall into 3 categories:

- reference dimensions, or pages ("Units of measure" of rank 1 and "Area" of rank 2, in the example): positions in these dimensions apply globally to the other two categories;
- stub dimensions ("Districts" of rank 1 and "Age groups" of rank 2, in the example): positions in these dimensions will appear as rows in the displayed matrix;
- header dimensions ("Gender" of rank 1 and "Years" of rank 2, in the example): positions in these dimensions will appear as columns in the displayed matrix.

33. The ranking number associated to a dimension in each category gives the embedding level of positions of the respective dimension in the extracted matrix layout.

34. Different perspectives of the same data can be obtained by interactively changing the matrix shape, that is moving dimensions from one category to another or changing the rank of the dimensions in one category.

35. Data extracted can be saved for later use and retrieved from previous savings.



		Gender	Years				
		Total		Male		Female	
Districts	Age group	Year 1980	Year 1985	Year 1980	Year 1985	Year 1980	Year 1985
Alba	Total	391686	403622	195563	202088	196123	201534
	10-14 years	36356	30286	18707	15569	17649	14717
	15-19 years	36510	35839	19047	18488	17463	17351
	20-24 years	27261	35321	14939	18621	12322	16700
	25-29 years	21741	25516	11166	13780	10575	11736
Arad	Total	492439	497143	238673	241330	253766	255813
	10-14 years	33540	23915	17032	12300	16508	11615
	15-19 years	37003	33990	18572	17120	18431	16870
	20-24 years	35120	38035	17900	18858	17220	19177
	25-29 years	26171	35374	13009	18044	13162	17330
Arges	Total	568854	607255	276868	297432	291986	309823
	10-14 years	48641	41459	24980	21273	23661	20186
	15-19 years	49854	48890	25549	25019	24305	23871
	20-24 years	43592	50406	21979	25749	21613	24657
	25-29 years	35341	43339	17912	22017	17429	21322

3.3 Metadata Display

36. Five categories of metadata for the current matrix are displayed:
- the indicator definition;
- time-related information (periodicity, start period, end period);
- methodology (incl. coverage, breaks, adjustment);
- data sources;
- other remarks.

3.4 Matrix Calculation

37. Two kinds of calculation can be performed with extracted matrices:

- calculations within the same matrix (sum, average, subtraction, division, percentage);
- operations using two matrices (join, merge).

38. Any of the calculations within one matrix (sum, average, subtract, division, percentage) is done along one of the matrix dimensions (the “control dimension”, defined by the user), according to certain rules. New calculated positions are added to the matrix, but they are available only to the local session, not in the database.

39. The definition for the matrix operations of join / merge is based on the existence of common / uncommon dimensions in the two participant matrices. A union / append matrix is generated, containing values from the two original matrices. In this way, matrices with common dimensions can be vizualized together. The result matrix can be further used by the user session in other operations, but it is not saved in the database, as it would be redundant.

3.5 Matrix Export

40. Any extracted matrix can be exported in an number of external formats (EXCEL, ASCII, HTML). Metadata are also exported if required.

3.6 Help

41. Help files are available on-line for the “Glossary of Terms” (explaining the notions used in the system) and for the “User Manual”.

4. Project Implementation

4.1 Technical

42. For performance testing, an amount of 75%-100% of the static database tables (having low update rate) and 10% of the dynamic database tables (matrix cells) were loaded into the target database. Tests results were accepted from a time response point of view; still, a few functions were identified for further improvement, using both server and client tuning means. The first version of the “Consultation” module, developed with a usual graphical user interface, was subsequently adapted to a Web interface.

43. The system was demonstrated to the NCS top management and statisticians using a development database. Although the demonstrations had a favourable reaction from the audience, an opinion was expressed that operating with terms like “matrix”, “dimension”, “positions”, “extraction specification”, etc. is quite understandable for the in-house statisticians, but it may be too complicated for an ordinary user. It was stated that external users will be assisted for the moment by the “Dissemination” unit inside NCS.

44. Also further enhancements were identified concerning the module functionality, related to matrix operations, to issuing graphics and maps, to matrix export in other external formats, etc.

45. Databases that are currently running in NCS will be converted to conform to the multi-dimensional principle and then migrated to the new TEMPO database.

4.2 Organizational

46. In order to put in practice the TEMPO database management and use on a regular basis, the following “actors” and tasks were established:

- the data administrator: coordinate the operations suitable for centralized performing, like data loading / updating, loading control, identification of system “weak points”;
- the domain responsible persons: define the database content for a subject-matter area (describing indicators, matrices, nomenclatures, etc., loading metadata, setting priorities for data loading, preparing files for data loading);
- the design / development team: develop new functionality of the system, perform database and application administration.

47. Different implementation stages and information flows were agreed between the actors. Major next steps in the system implementation are:

- revision of existing metadata;
- preparing data files for loading, according to the timetable and priorities agreed;
- testing the system performance with large volumes of data;
- tuning the system.