

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Geneva, Switzerland, 22-24 September 1999)

Topic (iv): The adaptation, evaluation and implementation of statistical metadata standards (terminology, taxonomy)

PROTOTYPE OF THE CLASSIFICATION SERVER IN THE CSO OF POLAND

Submitted by the Central Statistical Office of Poland¹

Contributed paper

I. INTRODUCTION

1. The aim of this paper is to share the experiences with the development of a prototype of statistical classification and nomenclature as a part of the whole statistical metainformation system developed in the Central Statistical Office of Poland (CSO). The database containing statistical classifications and nomenclatures is located on a server specially dedicated for this purpose (further "classification server").
2. Upgrading of the computer environment (hardware, network and software) in 1991-1997 introduced a new quality in statistical production. Transition to client/server architecture allows a two-tier architecture in the "classification server" prototype. Three-tier architecture using an Intranet solution is planned in the next version.

II. OBJECTIVES AND USERS OF THE CLASSIFICATION SERVER

3. The most important goal for the Polish CSO is to integrate the Polish statistical system on this server already implemented with that of the European Union (EU). Development of a classification server is moving in that direction. EU and other international classifications and nomenclatures which were already implemented on the server were fully integrated with equivalent Polish ones.
4. Special attention should be drawn to the metadata integration. Different sources of classifications, developed in the past for different goals, call for external (i.e. between classifications) integration. Internal hierarchy of data (i.e. within given classifications) requires integration on different levels of classifications. It is much easier to assure data integration with a computer system which allows on-line access to the data.
5. Another important objective was to rationalise users' access to different information sources of the statistics, especially to international and Polish classifications. In the present version, classifications are available in fast on-line mode to different groups of users, mainly statisticians. The next (Intranet) version will allow broader access within a wide area network. It is important that historical data be also available on-line. Interpretation of classifications sometimes has to be supported in on-line mode. This was done by introducing of a broad range of help messages and comments on different levels, precedence included.
6. Other important objectives are easier maintenance and dissemination of gathered metadata, documentation, integration with other statistical systems etc.

¹ Prepared by Stanislaw Sieluzycski.

7. The users of the "classification server" are statisticians involved directly in classification maintenance, those using classification data in other systems, especially in REGON - national business register on enterprises, statisticians responsible for dissemination of data, for publishing of data, external users from Tax offices, Custom Offices, Ministries, enterprises etc.

III. STRUCTURE OF THE DATABASE ON THE CLASSIFICATION SERVER

8. The prototype of the classification server comprises several economic classifications and nomenclatures. The following picture 1 illustrates relations between classifications considered. These classifications can be grouped as follows:

- (i) geographically:
 - world classifications: ISIC, CPC, HS, SITC,
 - European classifications: NACE, CPA, PRODCOM, CN,
 - Polish classifications: EKD/PKD, KGN, PKWiU, SWW, KU, PCN,
- (ii) according to statistical domains:
 - economic activities,
 - products,
 - foreign trade.

9. Other economic, social and territorial classifications and nomenclatures will be implemented and fully integrated with the above-mentioned classifications in the next version of the classification server.

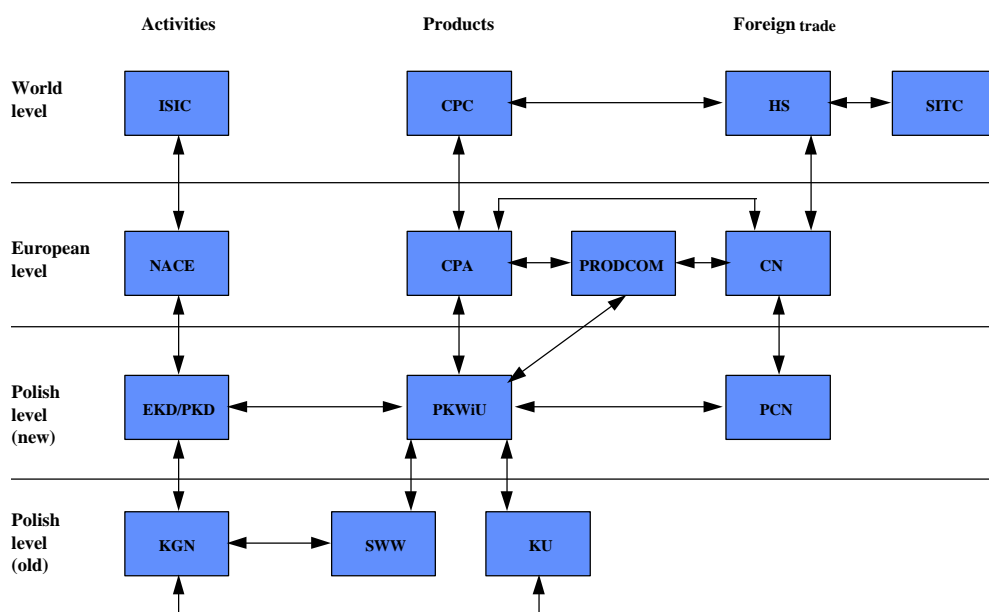
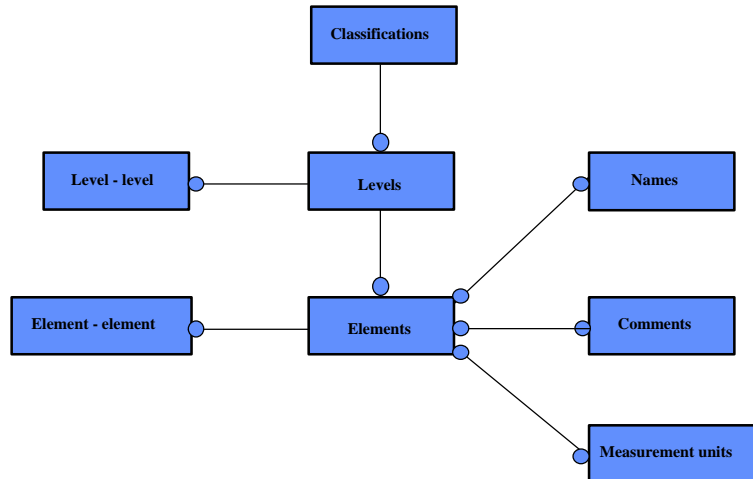


Fig. 1. Classifications schema

10. The general scheme of a database on classifications and nomenclatures is presented in picture 2. There is an assumption that each of the classifications can be defined on 3 tiers:

- (i) Classification tier - where the highest tier is a classification name, description of geographic and domain range, range of validity dates, comment on classification tier, author, classification manager, application area, classification rules. Each classification can have any number of levels (in practice 1 to 10);

- (ii) Level tier - this is a set of classification elements on a given (the same) detail level. For instance level: class in NACE classification is defined as a four digit number ("xx.xx"), while level: group has three digits ("xx.x"). Each level is identified within given classification and described by: level name, code format, range of validity dates, comment on levels tier. Each level have any number of elements (limited according to the code format);
- (iii) Element tier - the basic tier where classification elements are defined. Each element is usually of the same length for any given classifications level. Elements are identified within the given level of a given classification and described by: range of validity dates, element status, source and description of modifications (if any exist).



Pic. 2. General database model

11. Theoretically, the assumption of 3 tiers restricts mutual elements connection. Instead of a full graph or network, we have an ordered tree for each classification. In practice, however, we did not find a single classification where such assumption would lead to impossibility of implementation. This does facilitate the implementation significantly.

12. Classification elements (as well as classification levels) can create many relations of different types, such as:

- hierarchical relationship - where one element of a given level "N" has direct expansion to the elements of level "N+1". Such a relationship is typical and appears within a given classification;
- equivalent relationship - where one element of a given classification is equivalent to other elements of another classification. The term "equivalent" is not precisely defined (by rules) within classification. Instead, special transition tables (with transition keys) are defined by users involved in classification maintenance based on the meaning of both of the elements,
- covering relationship - where one element of a given classification has exactly the same meaning as the elements of another classification. A case when one classification is a subset of another one is typical for such relations.

13. Each classification element can have several names in different languages. The names can be printed

in a shortened or in full form. Sometimes CSO of Poland is not responsible for the naming of elements or naming conventions. In general, the relationship between elements and names are type of "N:N". It often leads to a time-consuming search process. In addition, the users often want to use the names for searching. Frequently used search techniques are as follows:

- simple search - which is time-consuming but easy to implement;
- search using keywords - which is time and human resources consuming during data preparation, because of Polish inflexion;
- search using thesaurus - which seems to be the best but is the most complicated to implement.

14. Each classification element could also have annotated comments and measurement units. Comments can be in different languages and of different type (precedence type is one of the most important ones).

IV. FUNCTIONS OF THE CLASSIFICATION SERVER

15. The following functions have been foreseen for the classification server:

- consultations – allow for searching of required information; searched data could be used directly, as a base for more advanced searching, then edited and exported to another system,
- history of modifications – allows for on-line access to historical data,
- upgrading – any modification of data can be done after verification of user rights and detailed specification of range of this modification,
- loading of metadata – useful function for loading of larger portions of data,
- reports – administration tool for diagnosis of server stage and history of users' access to the server (information on user connections to the database, realised data modifications, users rights, etc.),
- input to other systems,
- security management,
- information on classifications server.

V. TECHNOLOGICAL PARAMETERS

16. The classification server has been design in client/server architecture. The HP server runs under HP/UX operating system and Ingres RDBMS. Clients running under Windows 95 are connected to the server through ODBC to the database.