UDC 311.11

Tetyana I. Lumpova,
Candidate of economics

A Model for Metainformation Structure of a Statistical Indicator for Description of Statistical Observation

The article discusses a model for metainformation structure of a statistical indicator and selected aspects of its practical applications in integrated systems of data processing, managed by metadata, as a required condition for integration of indicators used in statistical observations in the single data repository. The author outlines the model components and their use in computerized data processing, with emphasis on data control. Application of the model is illustrated by the description of statistical indicator "Stimulating and compensatory payments".

Key words: *informative model, statistical metadata, statistical indicator, statistical information systems, statistical observation.*

Constructing a standard system of metadata for internal and external users is outlined as a foremost task in the Ukrainian Strategy of Official Statistics Development till 2017 [1]. To have this task fulfilled, the State Statistics Service of Ukraine has introduced Standardized Descriptions of Official Statistical Observations (SDOSO) as the core of official statistical observations [2; 3]. SDOSO information is designed for use when preparing the official statistical observations plan (OSO plan), compiling OSO descriptions for data users, and performing most part of production procedures related with OSO processing. An important potential purpose of SDOSO information is systematization of statistical indicators when compiling the catalogue for the Integrated System for Statistical Information Processing (ISSIP). Successful systematization of statistical indicators is based on determining the principles of organization and classification of statistical indicators as their systematized distribution by particular groups, classes, categories according to their similarity or difference, when each indicator is given with a particular code [4, p. 35–36]. In view of the accumulated information base of SD OSO, to perform such systematization, there arises the necessity to make the general formal description of the statistical indicator in the form of model which demonstrates its metainformation structure.

Such models for creation of the information systems (IS) are widely used in European statistics, in particular this issue was considered by K. Zeila [5] at Joint Sessions of the United Nations Economic Commission for Europe / Eurostat / Organization for Economic Cooperation and Development devoted to information technologies (MSIS) and by S. Bacelar [6] – at Joint Sessions devoted to statistical metadata (METIS), as well as by T. M. Isfan [7]. The works of Swedish scientist B. Sungren [8; 9] is a great contribution to formalization of the statistical metainformation description. This experience does not fully correspond to the approaches, traditional for the Ukrainian statistics, which are based on the principles stipulated by [10]. In particular it concerns the usage of the concept of "variable" in these models, as it is not equal to the traditional definition of the statistical indicator. There is no clear definition of the concept of "variable" in the contemporary national statistical literature. In scientific sources of the international statistics the concept of "variable" is usually used concerning the answers in the questionnaires and other forms of statistical inquiry, as well as in the descriptions of the IS components models, to describe a storage unit for statistical data. In [11] the variable is defined as "the characteristic of the observation unit, which can obtain more than one value from the set of values, which correspond to the numeral measure or the classification category (e.g. "income", "age", "weight" etc. and "activity", "industry", "decease" etc.)". Such an approach is based on the standard "ISO/ IEC 11179 Information Technology – Specification and Standardization of Data Elements". Another important aspect which should be considered in the model of the statistical indicator in terms of its use for ISSIP, is the need for a clear allocation of those components of the statistical indicator which are metainformation, whereby in the future it will be possible to make not only systematization and grouping, but also information search operations. It should be mentioned that the problems of metainformational structure of the statistical data are almost unexamined in the national statistics.

The purpose of the study is to determine the place of variable for the distinction between concepts of "variable" and "statistical indicator", and also identify the practical aspects of using the statistical indicator in IC, based on the model of metainformational structure made by the author.

Information collected for SDOSO purposes contains data important for the set of statistical indicators, for each statistical observation. Apart from the nomenclature of statistical indicators' names, it includes their definitions specifying their meaning, purpose or function. To analyze information about statistical indicators collected by SDOSO, a model for metainformation structure of a statistical indicator has been constructed and used to validate the adequacy of metadescription for the statistical indicators, given in SDOSO and based on descriptions

given the Glossary attached to the Plan of Statistical Observation, approved by the Decree of the State Statistics Service of Ukraine from 29.12.2009, № 498 (referred hereafter as Glossary) [2; 3].

Statistical indicator is defined in Glossary as "aggregate quantitative and qualitative characteristic of a phenomenon or a process – a statistical value that is computed, contrary to the parameters that are registered"; it follows that "the qualitative dimension of a statistical indicator reflects the meaning of a phenomenon or a process in the specific context of location and time, whereas its quantitative dimension reflects its size and absolute, relative or average value". Therefore, statistical indicators always show location and time of studied phenomena or processes and their measurement units. Statistical indicator's definition as the aggregate quantitative and qualitative characteristic of a phenomenon or a process provides the guideline in elaborating the framework for statistical indicator's metadescription covering the three components: basis (the first component), quantitative value (or variable, using the common terminology) (the second component), and attributes (the third component).

According to the Glossary definition, "the basis of a statistical indicator reflects the meaning, peculiar features and specifics of a phenomenon or a process, without specifying the conditions of time and location of a statistical observation and the quantitative value". It is, therefore, an abstract notion reflecting the meaning of characteristic of a phenomenon or a process studied in a statistical observation. Quantitative value of a statistical indicator corresponds to its quantity (size, volume or level), whereas attributes of a statistical indicator are a set of qualitative characteristics immanent to the basis of a statistical indicator, which, combined with the basis, makes the indicator unique. An example is the set of characteristics specific to area, time and classification [2; 3]. Attributes, therefore, constitute the set of parameters that clearly distinguish quantitative characteristics of an entity, a phenomenon or a process, defined by the basis, from the general set of values.

In view of the above-mentioned definition of a statistical indicator and its components that are also defined in Glossary, and considering the practices of Ukrainian [10, p. 33–34] and European [5–9] statistics, metainformation structure of a statistical indicator can be shown in schematic form (Fig. 1). It should be noted that not all the positions (cells) of statistical form / questionnaire can be described by the scheme given in Fig. 1. This, by far and large, concerns answers to questions in the questionnaire and cells of tables contained in statistical check form, which correspond to the category of classifications and, therefore, cannot be defined as a statistical indicator that is to be computed (such as average number of part-time employees) or fixed (such as the debt of physical persons for electricity supply as of the end of reporting month). Answers to the questions in questionnaire and the above-mentioned cells in tables contained in statistical check form are denoted by us as "registration parameters". Notions of "statistical indicator" and "registration parameter", therefore, need to be distinguished, with definition to be given to registration parameter.

Registration parameter can be defined as the characteristic of quality of an observation object, which does not reflect the quantitative value to be computed but fixes the quality of an entity that may have no quantitative characteristics in statistical check form or questionnaire. Thus, "Coupon for registration of residence location in Ukraine" is filled for each person and contains only this person's characteristics (gender, citizenship, country of origin, date of birth etc.).

Registration parameters belong to the category of classification, and they are transformed into the statistical indicator's attributes in computer processing. In the above given example, they characterize the indicator of number of persons that have come to Ukraine. Accordingly, registration parameter cannot have a measurement unit, formal characteristics etc. immanent to the quantitative value corresponding to a statistical indicator. The attributes of a statistical indicator, when combined with its basis, form an economically meaningful concept: they become a statistical indicator with the computed quantitative value (such as the total number of observation objects by selected registration parameters). Because a statistical indicator is going to be subject for further analysis, it should be noted that metainformation structure of a registration parameter requires separate definition, and the model shown in Fig. 1 cannot apply to it.

Code, name and definition of the basis of a statistical indicator, given in Fig.1, are combined in the component "Identification characteristics", because they, if taken in the totality, enable for a clear definition of the basis of a statistical indicator.

Formal characteristic of a statistical indicator (the component "Formal characteristic" in Fig. 1) is linked to its basis, because average and relative indicators are computed by various formulas, unlike absolute indicators that may be sums or aggregates of other absolute indicators, and it is indirectly linked to variable, its typology and type (the indirect link is denoted by thin stroked line). Formal characteristic "Other computed indicators" will be treated by us as the one immanent to coefficients, forecasting and integral indicators and other groups of indicators that cannot be strictly regarded as average or relative indicators.

By analogy, type of a statistical indicator (the component "Type" in Fig. 1) is linked to the basis and has indirect mutual link with formal characteristic, because additivity of a statistical indicator is largely dependent on formal characteristic. Also, type is indirectly linked to variable and, hence, to measurement unit and typology. These links are shown in Fig. 1.

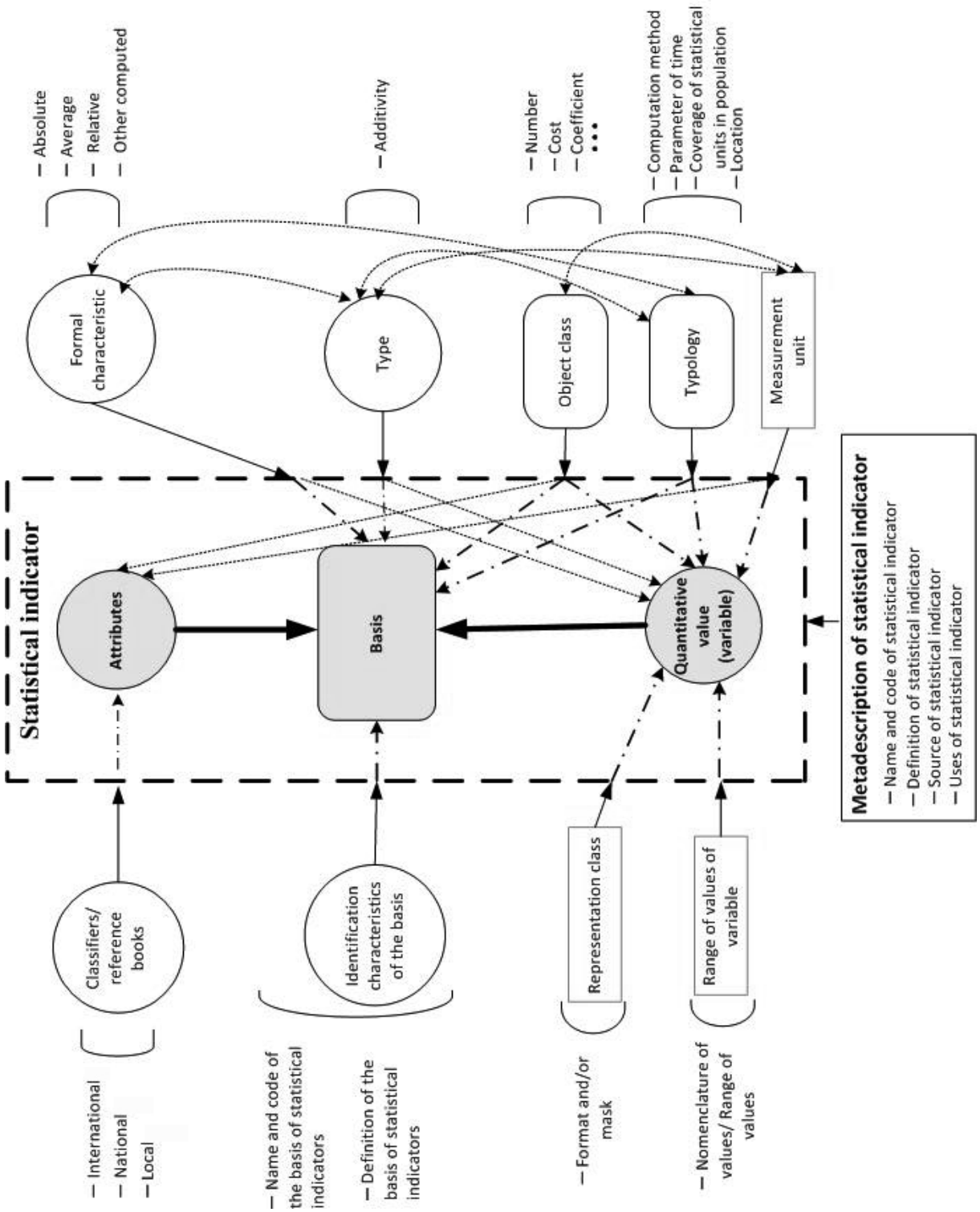

Fig. 1. Metainformation structure of a statistical indicator

The component "Object class" given in Fig. 1 is a set of ideas, abstracts or things of the real world, which predetermine properties and functionality (behavior by similar rules) for variables and, accordingly, for statistical indicators [7]. Thus, variables in the basis of statistical indicators "Salary fund" or "Aggregated salary debt" can be classified in one object class "Costs". In database (DB) management systems, object classes can be realized as domains constructed for specific purposes. Domains are more primitive types that are held in DB. DB technologies usually have a fixed set of domains, but in case of need in presenting the types that are not held in DB but based on its domains an applied mechanism for description of stable types of data can be constructed. This mechanism may include certain procedures for control, performed when a new value is embedded or an existing one is modified. Such domain can be specified for codes, in order to check the relevance of a new value to coding system. Given the definition contained in SDOSO, the following most common object classes can be outlined:

1. Number of persons.
2. Number of articles or objects, denoted by integers (pieces, agreements, units etc.).
3. Number as a measure unit, denoted by rational numbers, which can be further detailed by meaning (such as area, amount, volume, roominess etc.).
4. Cost as a measure linked to monetary unit of measurement.

Linkage of a statistical indicator to object class allows for setting up limitations on use of measurement units (thus, for absolute statistical cost value indicators, only monetary unit of measurement can apply), masks for quantitative value (for example, value only negative), range of values. This approach is instrumental in setting up analogous rules of control for a group of statistical indicators, instead of making analogous checks for selected statistical indicators which quantitative value belongs to one object class. Linkage of a statistical indicator to object class is shown in Fig. 1 through setting up the link (pointer).

Measurement unit (component "Measurement unit" in Fig.1) is linked to variable, because sometimes it depends on the specific value of an attribute, which is denoted in Fig. 1 through indirect link. Apart from this, measurement unit is linked indirectly to object class. Thus, measurement unit for the basis of statistical indicator "Retail sale per capita" depends on the meaning of an attribute describing a commodity in the Catalogue of commodities and commodity groups; so, it may be "liter", "kilogram" or "article".

Typology of a statistical indicator (component "Typology" in Fig. 1) is directly linked to its basis and quantitative value and indirectly linked to its type and formal characteristics, because the latter sets limitations on typology and vice versa; thus, average indicators cannot be estimated for a time moment, as they give average estimate for a time span.

Component "Representation class" identifies type of data or format and/or data mask for a variable and a statistical indicator, accordingly. Basically, type of data is defined as characteristic given to an object (variable, function, recording field, constant, data array etc.), either explicitly or implicitly. Type of data specifies a set of possible values, format for their holding, a size of allocated memory and a block of operations that can be made with the data. By data type, a statistical indicator can be integer or або real number.

Data format provides for the structure of data in files or DB tables, which depends on the specifications of information systems (IS) software and hardware. To compute quantitative value of statistical indicators and registration parameters at SDOSO level, it is enough to specify the number of symbols for numerical and text data, and the number of symbols in integer and fractional parts of real numbers.

Data mask is defined as "a block is symbols used for control of holding or withdrawal of selected parts of another block of symbols" [12], or as schematic mapping of data or template by which they are transferred to IS; thus, for a negative integer that is higher than –10, the mask "–X" can be set, with numbers from 1 to 9 being denoted as "X".

Component "Range of values for variable" specifies a block of permissible values for a given type of data and associated definitions for discrete or continual value. The definitions can be given in form of a range or a set of ranges and/or in form of a set of specific values, or, sometimes, in form of values which the variable cannot have. The link of data type with the range of values for variable can be illustrated in this way: the share measured in percents has data type "real number" with the set of values of all real numbers, and range of values of the statistical indicator's variable will range from 0 to 100.

Coding of properties (formal characteristic, typology, type etc.) of statistical indicators and other meaningful components the scheme shown in Fig. 1 supposes links with relevant national classifications, especially with the Classifier of the system of denotations for measurement and accounting units and local classifications (local reference books of formal characteristics, types etc.).

The below given Fig. 2 shows description of statistical indicator "Encouraging and compensation payments" using the scheme in Fig.1. For illustrative purposes, the description is given in simplified form.

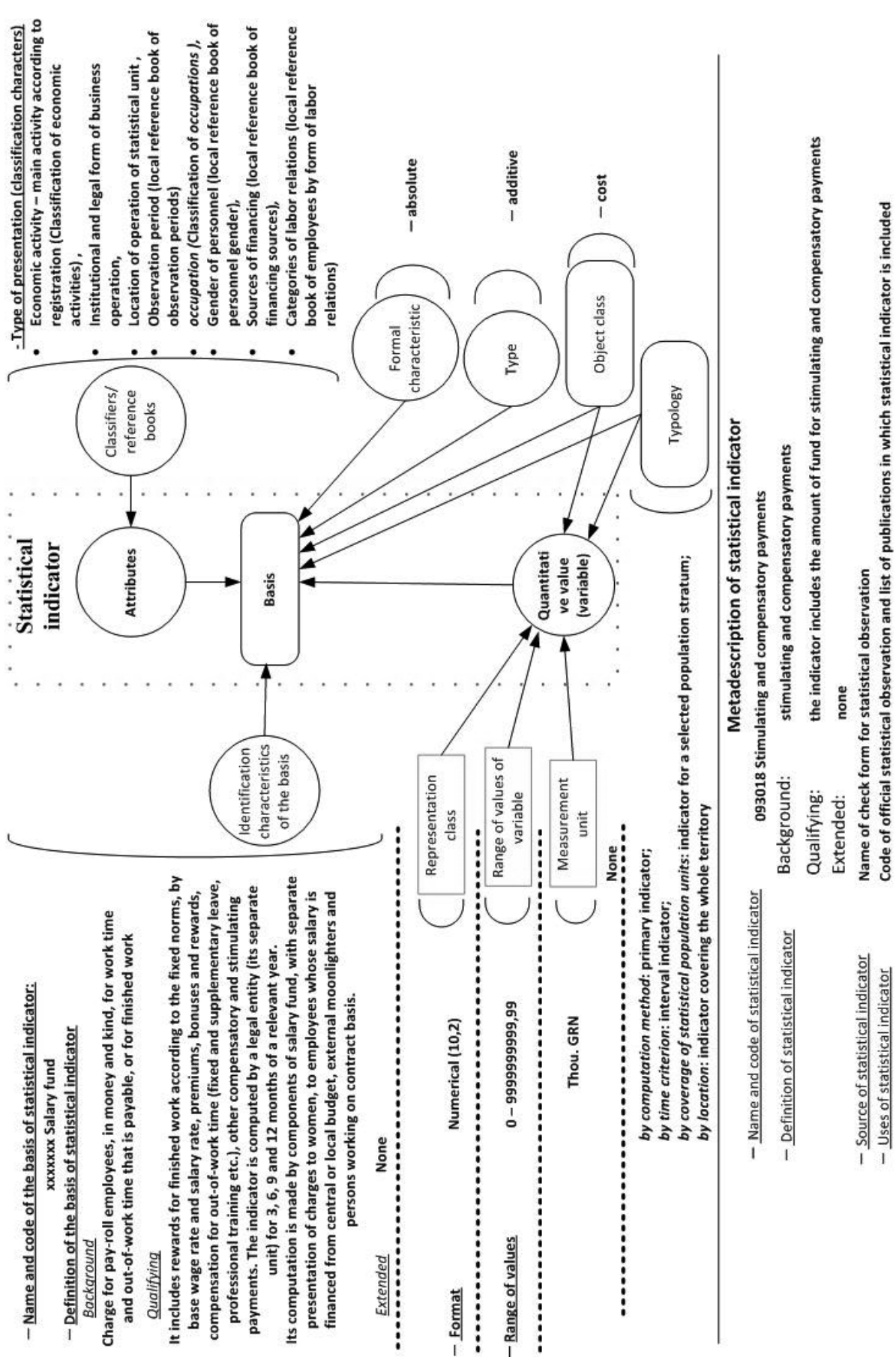

Fig. 2. Example of constructing meta-description of a statistical indicator

Practical applications of the model for a statistical indicator for description of statistical observation will allow for standardization of indicators' descriptions and better understanding of statistical indicator's components, which, accordingly, will enable for clarifying their definition. Standardized description of a statistical indicator is a required condition for integrating the information on OSO results in the single data repository, because this process requires constructing the integrated catalogue of statistical indicators. Besides that, identification of links between description components will allow for setting up control of the propriety of filled descriptions of statistical indicators when constructing tools to maintain metainformation base of statistical observations (for example, the catalogue of statistical indicators in the ISSIP framework), which can, accordingly, be used in computer processing of OSO data.

The proposed model can be a useful tool in constructing new statistical indicators and their characteristics. Apart from professional statisticians, the model is offered to specialists in information technologies, engaged in their introduction and maintenance in the statistics field.

The List of Used Sources

1. Постанова Кабінету Міністрів України "Про затвердження Стратегії розвитку державної статистики на період до 2017 року" від 20.03.2013 р. № 145-р [Електронний ресурс]. – Режим доступу : http://zakon4.rada.gov.ua/laws/show/145-2013-%D1%80

2. Наказ Державного комітету статистики України "Про затвердження уніфікованої форми опису державного статистичного спостереження" від 30.12.2009 р. № 505 [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua – Назва з титул. екрана.

3. Наказ Державного комітету статистики України "Про затвердження Глосарію до плану статистичного спостереження" від 29.12.2009 р. № 498 [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua – Назва з титул. екрана.

4. Парфенцева Н. О. Міжнародні класифікації в Україні: Впровадження й використання. / Парфенцева Н. О. – К. : Основи, 2000. – 351 с.

5. Zeila K. Metadata driven integrated statistical data management system [Electronic resource] / K. Zeila // Joint ECE/ Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS), Geneva, 17–19 May, 2004. – Access mode : http://www.unece.org/stats/documents/2004.05.msis.html

6. Bacelar S. Metadata Common Vocabulary: a journey from a glossary to an ontology of statistical metadata, and back [Electronic resource] / S. Bacelar // Joint UNECE / Eurostat / OECD work session on statistical metadata (METIS), Lisbon, 11–13 March, 2009. – Access mode : http://www.unece.org/stats/documents/ece/ces/ge.40/2009/mtg1/wp.15.e.pdf

7. Isfan T. M. Variables Subsystem [Electronic resource] / T. M. Isfan // Joint UNECE / Eurostat / OECD work session on statistical metadata (METIS), Lisbon, 11–13 March, 2009. – Access mode : http://www.unece.org/stats/documents/ece/ces/ge.40/2009/mtg1/wp.17.e.pdf

8. Sundgren B. An Information Systems Architecture For National and International Statistical Organizations [Electronic resource] / B. Sundgren // Meeting on the Management of Statistical Information Technology, Geneva, 15–17 Febr., 1999. – Access mode : http://www.unece.org/stats/documents/ces/ac.71/1999/4.e.pdf

9. Sundgren B. Metadata Systems in Statistical Production Processes – for Which Purposes Are They Needed, and How Can They Best Be Organized? [Electronic resource] / B. Sundgren // Joint UNECE / Eurostat / OECD work session on statistical metadata (METIS), Geneva, 9–11 Febr., 2004. – Access mode : http://www.unece.org/stats/documents/2004/02/metis/wp.3.e.pdf

10. Экономическая информация. Методологические проблемы / под ред. Е. Г. Ясина. – М. : Статистика, 1974. – 258 с.

11. UN Glossary of Classification Terms [Electronic resource] / Expert Group on International Economic and Social Classifications. – Access mode : http://unstats.un.org/unsd/class/family/glossary_short.asp

12. Системи оброблення інформації. Підготовлення і оброблення даних. Терміни та визначення : ДСТУ 2228-93. – [Чинний від 1994-07-01]. – К. : Держстандарт України, 1995. – 22 с. – (Національний стандарт України).

УДК 303.51:061.1(ЄС+ОЕСР)

Л. О. Ященко,
кандидат економічних наук,
завідувач відділу;
А. І. Шпеньович,
технік 1-ї категорії;
НТК статистичних досліджень

Формування вибіркових сукупностей для обстежень ділової активності підприємств у країнах ЄС та ОЕСР

Досліджено основні характеристики та етапи формування вибіркових сукупностей для обстежень ділової активності підприємств промисловості, будівництва, роздрібної торгівлі та сфери послуг у країнах ЄС та ОЕСР. Визначено ключові аспекти розрахунку розміру страт, побудови основи вибірки та панелі респондентів.

Ключові слова: *формування вибірки, стратифікація, розмір страт, обсяг вибірки, обстеження ділової активності підприємств, пропорційний розподіл, розподіл Неймана, статистичний реєстр, бізнес-реєстр.*

Під час побудови вибірки неминуче виникають проблеми, пов'язані з визначенням одиниць у генеральній сукупності, різноманітністю видів діяльності у рамках одного підприємства, створенням вертикальних і горизонтальних об'єднань. На жаль, у різних країнах усі ці питання вирішуються по-різному, використовуються власні визначення, що ускладнює формулювання загальних рекомендацій. І все ж таки кожна країна повинна мати максимально чітку інформацію про вид діяльності, розміри й розташування підприємств.

Метою статті є дослідження досвіду формування вибіркових сукупностей обстежень ділової активності підприємств (далі – ОДАП) у країнах Європейського Союзу (далі – ЄС) та Організації економічного співробітництва та розвитку (далі – ОЕСР).

Основні положення щодо формування вибіркової сукупності ОДАП викладені в Об'єднаній гармонізованій програмі ЄС щодо обстежень тенденцій ділової активності бізнесу та споживання [1] і у Керівництві ОЕСР із обстежень ділової активності [2]. Дослідженням зазначених проблем у країнах ЄС та ОЕСР займаються Дж. Пеллісьє (G. M. Pellissier) і Д. Нел (D. G. Nel) [3], М. Пугачова [4], С. Цухло [5] та ін.

За рекомендаціями ОЕСР, щоб побудувати коректну вибірку для цілей ОДАП, необхідно для кожного підприємства мати три ознаки: вид діяльності за класифікацією видів економічної діяльності (NACE), розмір залежно від кількості працівників (або інших показників, таких, наприклад, як оборот продукції, обсяг випуску, обсяг реалізації) і місце розташування (наприклад, у регіональному розрізі) [6].

У Керівництві ОЕСР рекомендовано три способи організації збирання даних у ході ОДАП [2]. Перший – найпростіший – передбачає опитування всіх підприємств. Це дорогий і тривалий метод, якщо тільки попередньо не була визначена невелика цільова сукупність підприємств для обстеження. Як зазначено у Керівництві [2], такий варіант є небажаним для ОДАП. Його можна реалізувати тільки у невеликих країнах Європи з високим рівнем розвитку комунікацій.

Другий варіант організації обстеження здійснюється за допомогою цільового відбору підприємств із генеральної сукупності. Однак такий спосіб, на думку експертів ОЕСР, теж небажаний для ОДАП.

Найбільш прийнятним є випадковий відбір із установленою ймовірністю включення (відбору) для всіх підприємств вихідної сукупності. У цьому випадку не потрібно ніяких припущень про репрезентативність, за випадковою вибіркою можна оцінити параметри генеральної сукупності та їхню точність. Саме такий спосіб формування вибірки й рекомендується використовувати для ОДАП. Випадкова вибірка може бути організована різними способами, а її ефективність залежить від обсягу інформації про підприємства.

Для визначення обсягу вибірки за різними стратами є два важливих критерії. Перший – це зручність, тобто вибирається спосіб пропорційного розподілу. Другим критерієм є точність, що зумовлює вибір оптимального розподілу. Там, де витрати на формування вибірки з різних страт однакові, оптимальна формула розподілу називається розподілом Неймана. Більшість країн – членів ОЕСР для визначення розміру страт використовують пропорційний розподіл, хоча деякі країни, зокрема Італія та Велика Британія, з цією метою застосовують розподіл Неймана. Обсяги вибірок зазвичай залежать від розміру країни, але національні організатори опитувань гарантують їхню репрезентативність і можливість виконання запитів користувачів результатів.

У теперішній час країнами – членами ЄС щомісячно проводиться чотири обстеження підприємств у таких галузях, як промисловість, будівництво, роздрібна торгівля та послуги. Деякі додаткові обстеження щоквартально здійснюються у промисловості та будівництві. Крім того, двічі на рік відбувається інвестиційне обстеження промисловості, у ході якого збирають інформацію про