

Heikki Rouhuvirta
Statistical Methodology R&D
Statistics Finland¹
FIN-00022 Statistics Finland
heikki.rouhuvirta@stat.fi

For CODACMOS project
Bratislava 7th October 2004

An alternative approach to metadata – CoSSI and modelling of metadata

1. As an introduction

I present my basic ideas about statistical information as certain kinds of axioms. It is not possible to analyse these axioms closely in this context so I will not elaborate on them here, but they constitute my points of departure in selecting the means and methods utilised in the solution:

1. The question concerning reality; the task of statistical data processing is not to model reality or its processes. The modelling of reality has been done by the statistical researcher and the outcome from the modelling has been implemented in a statistical survey with its own methods. The task of statistical data processing is only to arrange, organise and process as efficiently as possible the data from a statistical survey for a statistical analysis. Thus, we do not need to analyse different processes of reality or methods for the modelling of reality when we try to find more efficient models for processing statistical data.
2. The question concerning the nature of statistical data; statistical data are already fully defined when they are created. No units or elements of statistical data are collected until the statistical data have been conceptually and operationally defined.
3. The question concerning the structure of statistical data; conceptual defining of statistical data produces for them a fully defined logical structure that is characteristic of them. Thus, the processing or dissemination of statistical data does not need to resort to methods that have been developed to describe e.g. semi-structured data, or to producing structured descriptions from such data. Fully optimised methods and technologies developed for the processing of completely structured and exhaustively defined data can be exploited direct for statistical data.

If statistical information fails to meet these criteria or is in some other format, we have at some stage of data processing damaged or wiped out, destroyed or lost some information.

Bearing in mind these points of departure we set out to develop the CoSSI definition with the objective of organising statistical data so that they also contain statistical metadata.

¹ The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Finland.

2. Statistical production process – first approach to metadata

Barring a few exceptions, the process of producing statistics has, as a rule, been depicted during the past couple of decades as a production chain stretching from the collection to the dissemination of data. The “standard” components of the chain are the data collection phase, the production phase and the dissemination phase.

A part of the statistical production process that is significant from the point of statistical information is left outside the description and, thus also, outside the examination as a whole. The part concerned is the phase during which the topic the statistics depict and the contents of the statistics are defined. This phase is essential because it produces most of the metadata for statistics. The metadata produced during this phase are utilised in a number of contexts during the other stages of production. When statistics are published, some of the metadata are attached to them, either as direct quotes or annotations. Through these the user of the statistics receives information on e.g. the used international classification standards and their pertinent specific national application rules.

This “trimming” of the statistical production process has had certain aggravating, one could almost say fatal, consequences. The “overlooking” of the first, genuine phase of the production process has meant that no scope has been given for metadata within statistical production systems themselves. Different procedures for processing metadata and actual data have been adopted in these systems and combining them with the available technology has proven problematic and laborious as far as the design, implementation and maintenance of the systems, as well as the required statistical work are concerned. Additional work is required in adjusting the metadata into system-compatible contents and formats and in re-entering the metadata more or less manually into different systems serving dissemination.

The motive for the present study was the desire to eliminate the productional disparity between metadata and actual data, and its side effects. In an effort to do this, statistical information must be examined as a complete entity. Answers must be sought to questions such as what is meant by statistical information, does it have a certain structure or form and, if so, can they be universally defined? In seeking these answers one has to bear in mind that the principles of statistical production stem from the empirical research tradition and from the statistical science, not from data processing and its inherent, urgent questions.

3. Interpretation of statistical figures – second approach to metadata

In the data collection situation, the data provider uses some means to supply the data requested and specified by the producer of statistics. When the published statistics become available to the data provider, he or she also becomes a user of statistical information. In simplified terms, this means that anyone can use an interface to supply data, as well as use data with identical contents to those of the supplied data, albeit in processed form. In recent years, by far the biggest transformer of the environment in which public statistical information is used has been the Internet. As a matter of fact, this transformation is even greater than the one that was brought about by the development of databases. It would be reasonable to assume that in terms of their meaning the contents of the data would remain unchanged from the start of the

above-described process, i.e. specification by the data producer, right up to its end, i.e. use of the data, but is this the situation in reality?

We exclude from our examination the individual and social aspects that associate with meaning and human understanding, and focus on the informational and technological components of the process. In an ideal process, the meaning of the data does not change unless it has purposely been altered during data processing. In practice, however, data processing interferes with the meaning.

The main means by which the effects from a disturbing element can be minimised are improvement of the manageability and handling of metadata. This perspective is adopted in analysing the characteristics that are needed from the used technology and in evaluating the suitability of the diverse available technological solutions for the handling of statistical data.

An interference means that the meaning of data changes between their specification and their use, that is, from when the data leave the interface to when they return to it². This is a question of data quality, for the meaning of the data should not change during the process and the only thing that should change during data processing is the degree of their processing. In compliance with this, I refer with interference to an identified amendment of meaning between the specification of data and the use of data. If the meaning of data has altered during the process, interference has taken place.

In simple terms, the interference produced by diverse processing stages can be divided into two components, of which one relates to the understandability of data and the other to the appropriateness of data processing³. Essentially, both of these concern metadata. The understandability of data is determined by data descriptions, and the effect of data processing on the meaning of data can only be verified against processing descriptions and produced control data. The superiority or inferiority of a used system is largely determined by its capability to handle these descriptive data.

On the other hand, one may well ask what kind of an information technology system would best support the handling of descriptive data and what this support should then be like. Many factors associated with the organisation of the process may increase the vulnerability to errors. Likewise, many factors associated with information technology may increase the possibility of errors. An example of this would be a complex data system that does not support the documentation of processing or that allows or actually demands the use of complicated data structures, which in turn would increase the complexity of the processing procedures. In this respect, relational database solutions have represented retrograde development in statistical data processing if they have been used for storing or processing of statistical data. This is

² In the case of electronic administrative services, the data are always "born" in the interface even if the statistician did have an interface of his or her own that is separate from the interface meant here. The interface is a forum for publishing the specification of the data where a question is presented in its final format for the first time. This final formulation of a question is also the real specification of the data.

³ In processing, the influencing factors can be itemised more closely than presented here, whereby the factors affecting the meaning and the understanding of the meaning of data would be:

- 1) Description of data
- 2) Description of processing
- 3) Freedom from errors in processing

However, we do not discuss here any incidental or systematic errors that occur in processing and change the meaning of data from the intended.

the case if they are compared to, for example, solutions based on sequential files and a simple processing language.

However, in this context we are more interested in the share of the descriptive data and in how the effects from noise or interference can be reduced by solutions associated with descriptive data. This is justified because the range of tools offered by the database solutions used in statistics production for the handling of descriptive data is actually very limited.

In statistical information, descriptive data are textual information attached to figures in diverse ways. In the statistical process this descriptive, text-format information can be divided into the following types of metadata with different characteristics and properties:

1. Content-specific statistical metadata necessary for the interpretation of statistical figures.
2. Metadata connected with the identification and archiving of data, forming the metadata concerning a document.
3. Metadata concerning processing, of which some relate to statistical and methodological processing and belong to statistical metadata while some are technical metadata required by applications and belong to the process description.
4. Technical metadata concerning a process, which comprise the technical data required in applications and the metadata used or created in process control.

The system solutions that have been used for handling metadata in the production and dissemination of statistics have been various relational databases, in which metadata have been saved alongside numerical data into the same database or to an entirely separate relational database that has formed a metadata repository. The typical characteristics of these relational database solutions have been, on the one hand, general brevity of metadata and a large relative proportion of technical metadata required by the relational database solution of all available metadata and, on the other hand, laborious maintenance of metadata and difficulty in making metadata available for use in the dissemination of statistical information. In practice, the largest and most important proportion of the metadata required in the processing and dissemination of statistical data has been external to the information systems used in the process, actually manually managed by people.

In fact, it is impossible to describe statistical metadata exhaustively with the relations of relational algebra. This becomes quite clear when we examine the nature of statistical data, the contents of statistical metadata and the complexity of metadata.

4. Statistics as empirical data – third approach to metadata

Any evaluations of the suitability of a certain technology for the processing of statistical data should set out from a reflection on what kind of information statistical data actually are.

First, statistical data can be defined as empirical data bearing all their inherent hallmarks are⁴. The basic unit in empirical data, as well as in statistical data, is an ob-

⁴ The term statistical survey is here restricted to those statistical studies that produce descriptive or analytical information about society for the needs of social or economic decision-making, scientific research or international comparisons. Their aim is to describe the state of the whole population with re-

ervation. The actual interest focuses on the characteristics of the observation units that are being measurable in diverse ways. In statistics production, the observation unit is often referred to as the statistical unit.

Second, empirical statistical data are always static, cross-sectional information on the basis of which the previous state of the observation unit cannot be restored – neither is it the intention to do so. Cross-sectional data can be used to form an idea of the change that has taken place temporally, but this is done by comparing statistical information relating to two different points of time.

The third characteristic that defines statistical data is that as empirical data they must be organised in a particular, and maybe even highly specific way before they can be analysed. The data must be organised before they can be examined with empirical analysing methods. The most simplified form of organising is an observation matrix, which is also the basic format of statistical information. Even if statistical data were organised otherwise, into some other format, they must be managed in accordance with the observation matrix logic when they are being descriptively summarised or otherwise analysed.

As statistical data have thus been defined as empirical data, it would be justifiable to presume that the data concerning an event would also be empirical and this is, in fact, the case. In data concerning an event the observation unit is the event itself, and the data registered on the event in the characteristics vector. However, as the need to manage data is primarily dictated by the necessity to record them and the interest rarely lies in actual analysing of a set of events, a method has been developed for organising event data that is optimal from the point of their manageability, i.e. registration. The method breaks down data concerning events into more easily managed components.

If the desire is to draw generalised conclusions from event data in respect of one of their components, they must be aggregated relative to the observation unit that is to be formed. This produces one or several observation matrixes per each examined component. Data warehouse solutions contain examples of how the aggregation can be performed in practice.

4.1. Metadata of empirical data

In addition to the fact that the identification and physical recording of statistical data require their own descriptive data, statistical data have two dimensions, one of which identifies the observation unit and the other the measured characteristic of the observation unit. A description attached to the identification code of how it has been formed would suffice as the simplest metadata relating to the identification of an observation unit. The situation is considerably more complex with metadata that describe data on the characteristics of an observation unit. As a matter of fact, in practice, these metadata determine totally the information content of data.

In empirical research methodology, describing of the characteristics of information content always first requires a definition of the applicable concepts, usually a written one, then operational defining of the concepts, also often done in writing, and thirdly a description of the method for measuring the operationally defined concepts specifying the used scale, classification and measurement unit. These metadata as-

spect to the subject matter under examination (see Handbook on Quality Guidelines for Official Statistics, Statistics Finland 2002, p. 27).

sociated with measured characteristics are described, and must be described, before embarking on the collection of empirical data .

Once collected, data are edited in one way or another. As a result of the editing the meaning of the collected data may alter, whereby new metadata describing the data are created. The editing is usually targeted separately and in different forms at certain measured characteristic variables, whereby the new metadata must be attached to the descriptions of these specifically identified characteristics. This must be done, for example, when the classification applying to a variable is changed.

The methodological information that is created in the summarising and analysing of data, and which is also essential for their interpretation, brings its own addition to metadata. Summarising of data by e.g. cross-tabulating a number of variables produces a result where the variable structures cannot be subsequently altered without altering the result at the same time. It is, therefore, essential for the interpretation of the result that these structures can be presented and described to the users of the data as they appear in the original result, and in a format the user can understand.

5. Metadata system of statistical information⁵ – certain basic requirements

The characteristics of the metadata of empirical data as described above are also valid in describing the contents of statistical data. The metadata of both statistical and empirical data are characterised by a clear descriptive structure and logic. In compliance with this logic, descriptions of data content are organised so that sequential definitions clarify each other in succession. This imposes the following requirements on a metadata system of statistical data:

Requirement 1: The used metadata system should be capable of preserving and describing the structure and logic of statistical metadata.

On the other hand, data editing and processing create new metadata, which it should be possible to attach to the existing metadata to clarify them further. This leads to the following second requirement on a metadata system:

Requirement 2: It should be possible to use a metadata system to record accumulative metadata and to store completely new and more precise descriptions without breaking up the existing structure or logic of metadata.

With regard to the controllability of data processing the above imposes on a metadata system the following:

Requirement 3: The metadata in a metadata system must be available at all stages of the process for the purpose of verifying the meaning of data.

⁵ A metadata system is here understood as an abstract system and not as a concrete application.

The archiving of statistical data and the preservation of the meaning of statistical data impose on a metadata system the following:

Requirement 4: It must be possible to attach metadata to the data being processed at any given time, or to an output produced from them, most frequently a table, and statistical metadata cannot exist as separate, general metadata as which it would be of no use from the point of data processing and interpretation of printed out data.

6. CoSSI and statistical metadata

To clarify the problematics we have been developing a common structural definition of statistical information (CoSSI⁶), which covers different ways of statistical data organization (statistical data matrix and statistical table) and within which the structuring of the metadata connected to statistical data can also be implemented.

The point of departure in the CoSSI (Common Structure of Statistical Information) was an (infological) analysis of the information being considered. The conclusion from the analysis was that although in practice the definition of statistical information has varied according to a given situation and application, in reality, statistical information has a certain simplifiable and acceptable universal structure. The CoSSI describes the general structure that is not dependent on a situation of statistical information presented in differing formats.

The defining of the structure was not restricted in advance by selecting or specifying a certain application technology, which would have automatically determined or limited the volume or properties of the information that was to be analysed. The same also applied to the choice of the method used for describing the information, for it can be quite fatal if the applied technology requires that certain limitations or simplifications irrelevant to the information be included in the model. In fact, such limitations and simplifications narrow the content of the information being considered, and may even cause outright loss of information. On the other hand, the demands imposed on the used description technology must not be excessive, either. It is sufficient for the used description technology to meet the minimum criteria necessary for the presentation of results from an analysis of the information.

Minimum requirements were set on the method for describing statistical data, the most important of which were:

- 1) It should facilitate the presentation of statistical data as hierarchical information, and
- 2) It should facilitate description of organised strings

The hierarchy of statistical data is caused by three factors:

⁶ More detailed description is presented in Rouhuvirta and Lehtinen, Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003.

1. Nested variables form a hierarchical structure,
2. Class values formed for the classification of values received by statistical variables are often hierarchically arranged and in standard classifications form the hierarchical structure of the classification,
3. Statistical metadata are hierarchical in themselves; a characteristic connected with an observation object is first defined conceptually, from which a definition of measurement is then derived (so-called operational definition of a variable).

In addition, statistical metadata are largely presumed to be textual. This is also the case when statistical metadata are presented with conceptual symbols as a formula. Our cultural environment requires metadata to be multilingual. Within the CoSSI this has been solved as part of the structuring of statistical information, but justifications for the bases of multilingual solutions are not presented or argued here.

Setting out from these points of departure, in compliance with the CoSSI the structure of statistical metadata has been presented in simplified form as a tree structure using the ELM technique⁷ in Figures 1 and 2. The description of statistical metadata does not contain information about the processes that guide the production of statistics or about the monitoring of this process, or about the technical descriptions of data that are required by the diverse software programs used in the processing of statistical data. The document identification and other metadata required in archiving have been described in the other component of the CoSSI that covers metadata concerning documents and file copies. However, this part is not discussed in detail here.

⁷ About a tree structure using the ELM technique – see e.g. Maler, E. & El Andaloussi, J. (1996) Developing SGML DTDs. From text to model to markup. Upper Saddle River (NJ), USA: Prentice Hall, appendix B.

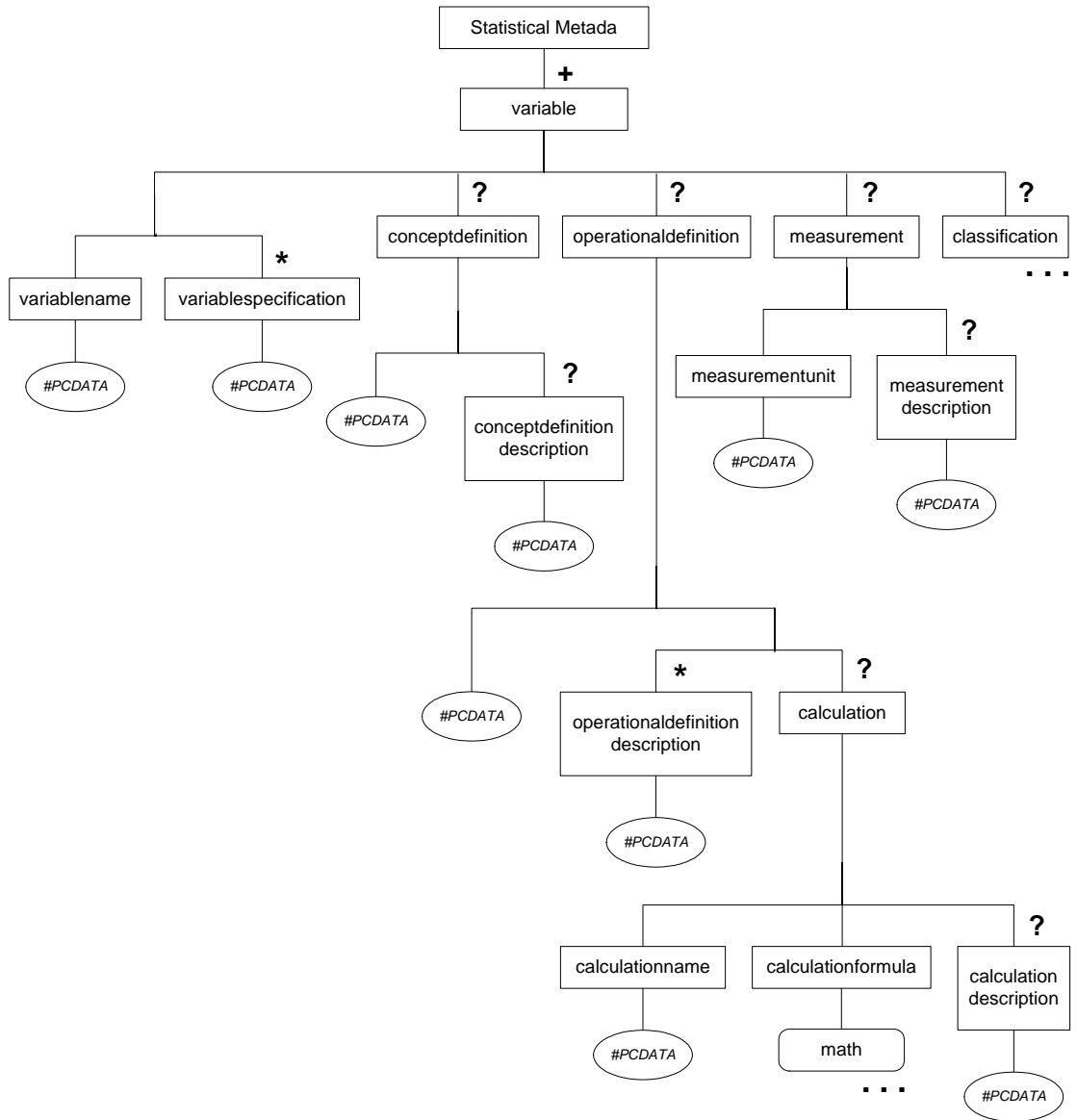


Figure 1. Logical structure of statistical metadata

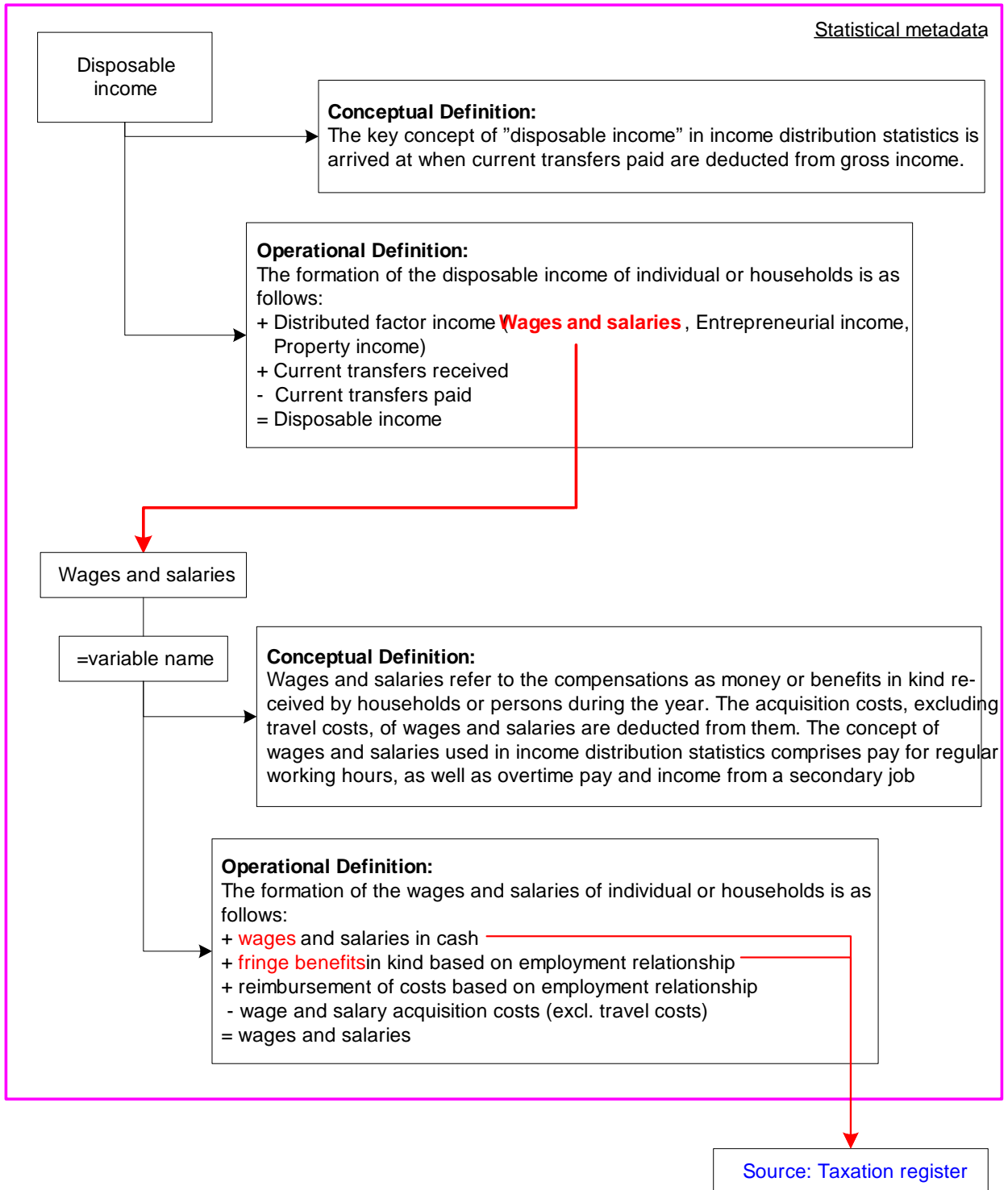


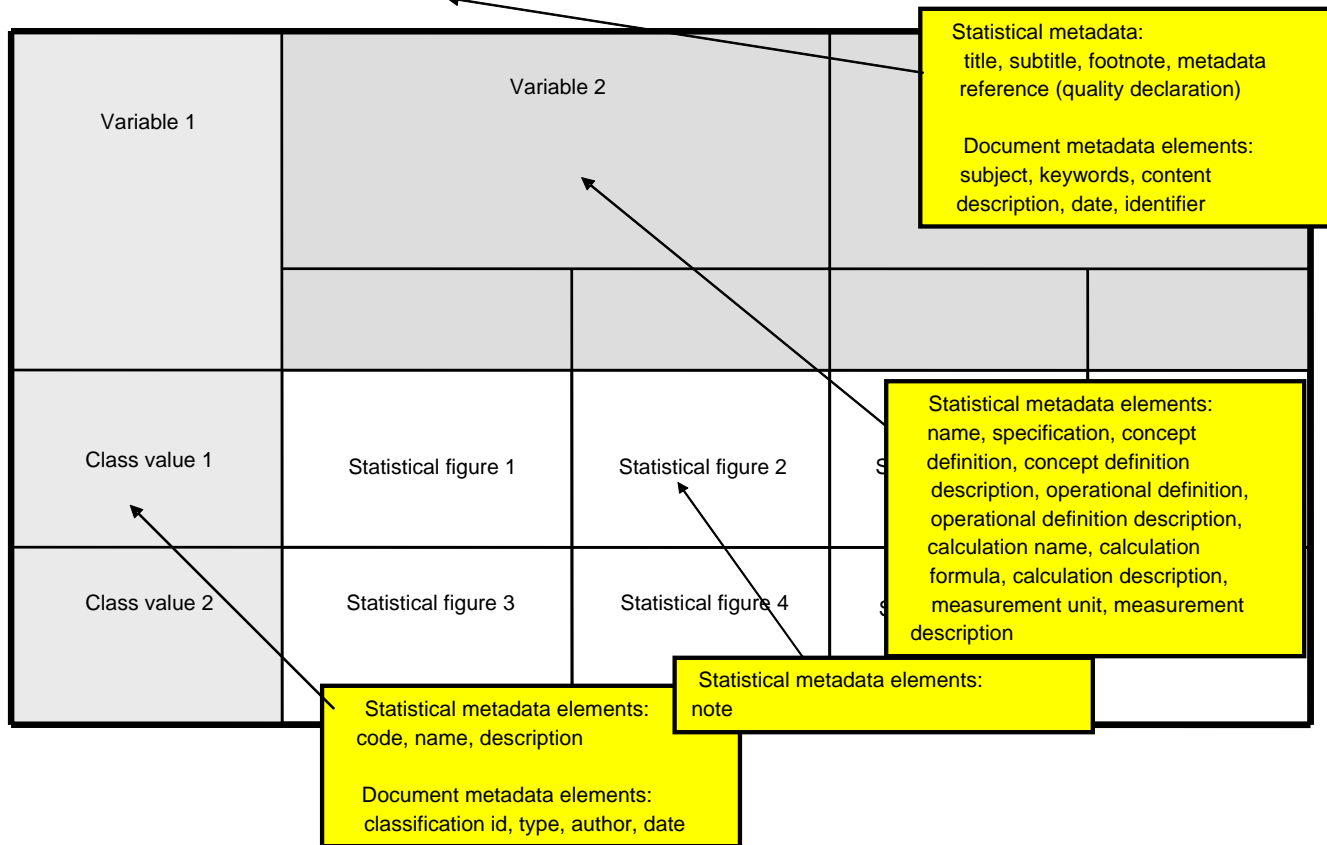
Figure 3. Metadata of Income Distribution Statistics according to CoSSI metadata specification

6.2 Informative table based on CoSSI metadata

From the perspective of users of statistical information novel content of metadata based on CoSSI metadata specification can be illustrated with the following table

description (see. Table 1.⁸), which lists all the metadata obtainable for users of statistical information.

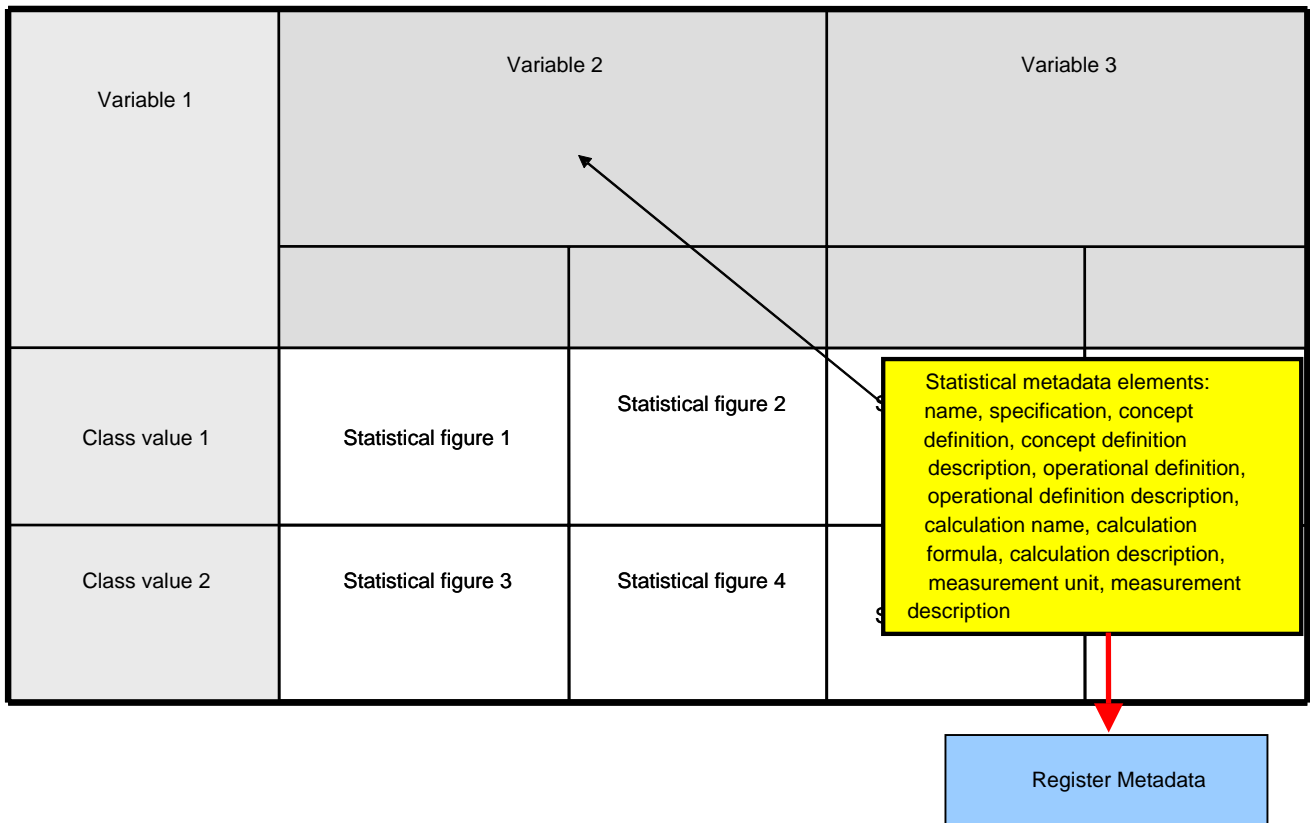
Table 1. Statistical metadata in a statistical table according to CoSSI



In this case users of statistical information are able to evaluate the usefulness of the statistical data and to choose the correct way of using it on the basis of almost all the relevant statistical metadata because users are able to access the statistical metadata from the very beginning of adoption of statistical information.

However, in this situation we still have a problem how to show the metadata of original sources of statistical information. The problem could be illustrated for instance with administrative register data. When in statistics exist derived variables based on administrative register data, the users should be able to obtain the register metadata as well (see Table 2.).

⁸ Ranta (2004), Statistical metadata - how far (or close) have we got. European Conference on Quality and Methodology in Official Statistics (Q2004).Mainz 24–26 May 2004.

Table 2. Register metadata in a statistical table

In Codacmos Project we tested the management of administrative metadata by CoSSI metadata specification.

6.3. An experiment of implementation/utilization: Taxation metadata in statistics

The most important question concerning methodological development is connected with the generalisation of a methodology: is CoSSI methodology usable for constructing metadata to describe the contents of administrative data. Data from administrative registers, such as those on taxation, are exploited extensively in statistics production in the Nordic Countries. Thereby the question concerns whether the same model can be used for metadata to describe the contents of taxation data or whether metadata concerning the contents of taxation data can be introduced into statistics production at the same time and using the same means as when data are being transferred, and then used as statistical metadata in statistics production.

The starting point to Taxmeta demonstration in Codacmos⁹ was to analyse a logical structure of taxation concepts. The main results of this analyse could be presented in following tree diagrams (see figure 4. and 5.).

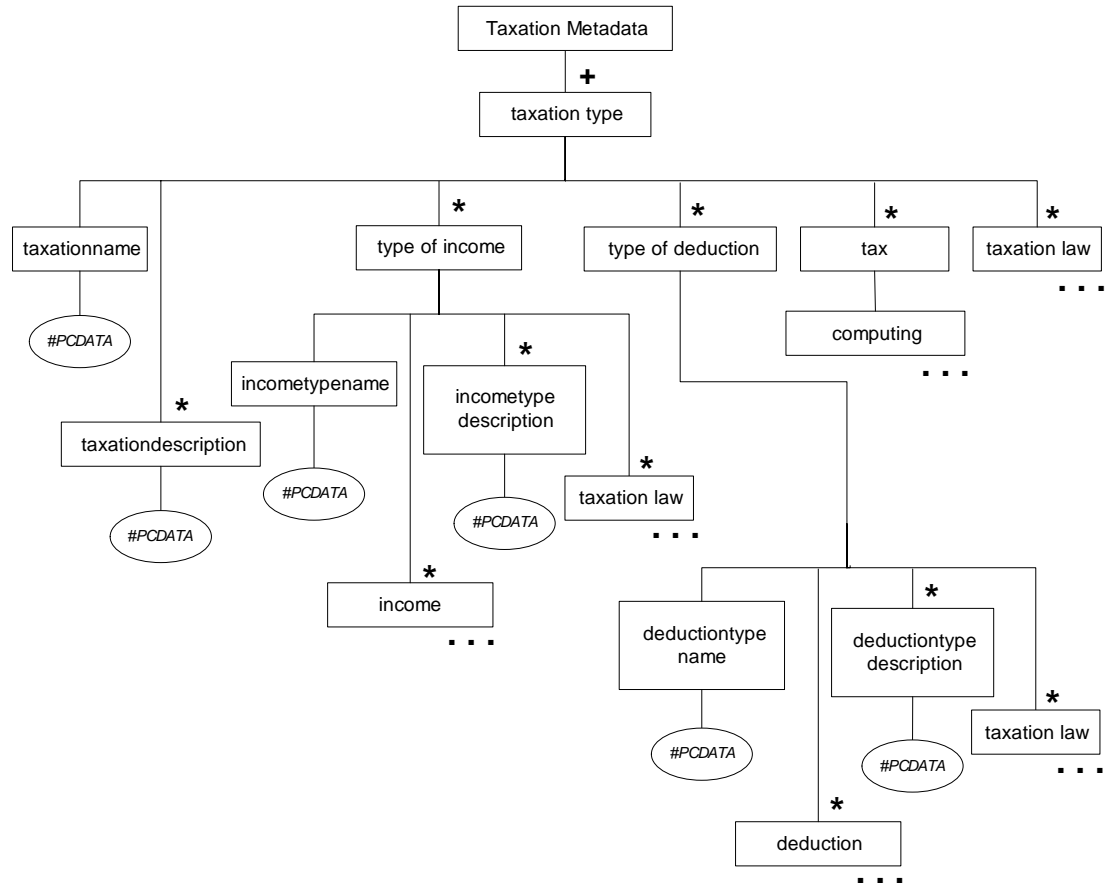


Figure 4. Logical structure of taxation metadata

⁹ See Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., (2004), Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos Project IST-2001-38636.

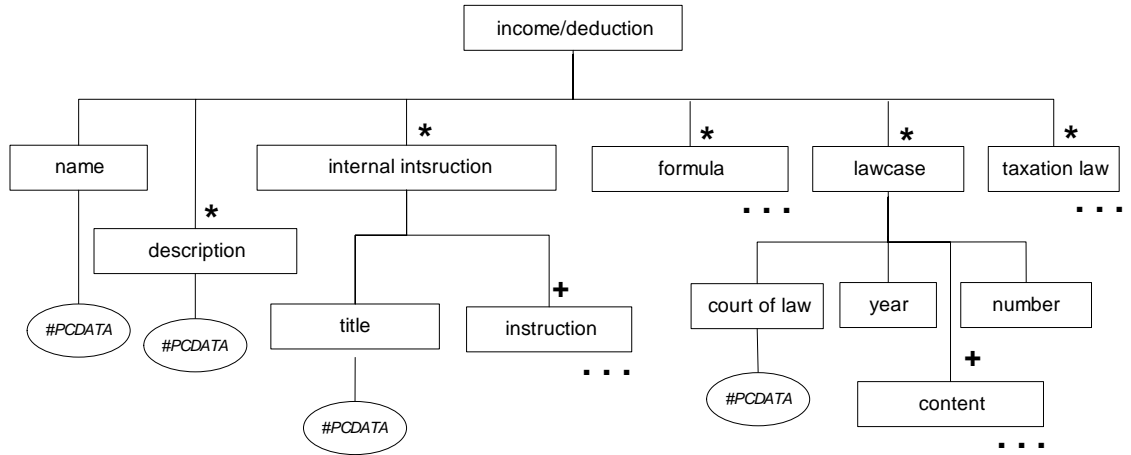
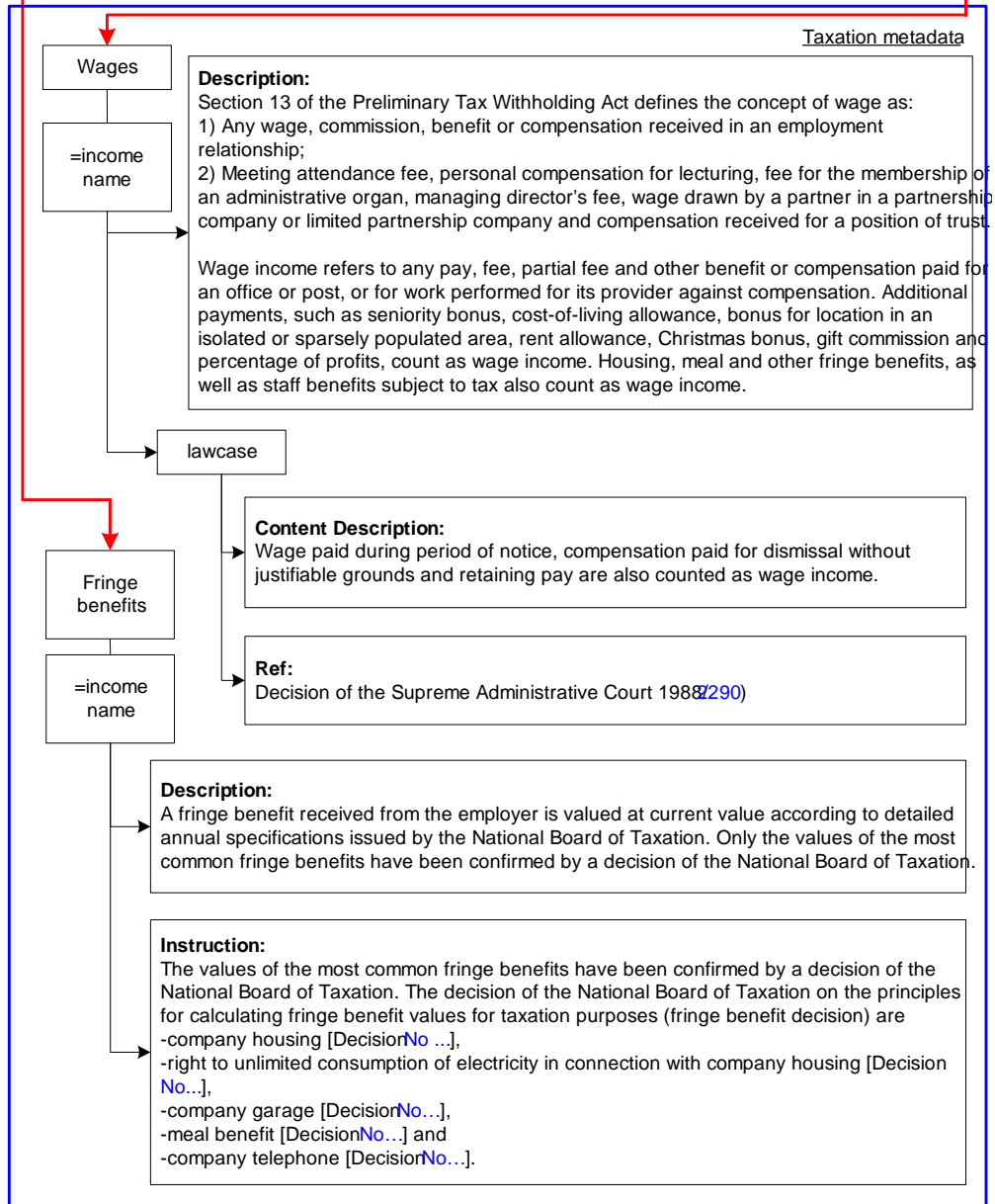
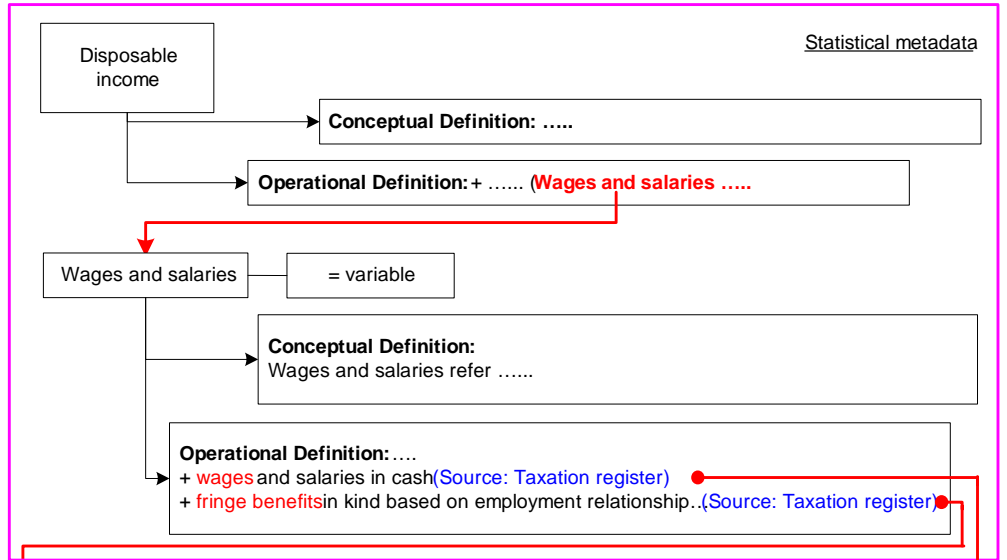


Figure 5. Logical structure of statistical metadata – incomes and deductions

When in this way structured taxation metadata is combined, for instance, with above presented metadata example of Income Distribution Statistics, it is possible to automatically generate the following metadata for statistical purposes:



All described metadata information could be obtainable to the compiler of statistics as well as to users of statistical information, for instance by Internet.

References

Laiho, J. and Hietaniemi, L. (Ed.), Quality Guidelines for Official Statistics, Statistics Finland, Handbooks 43b, Helsinki, 2002.

http://www.stat.fi/tk/tt/laatuutilastoissa/cont_en.html.

Maler, E. & El Andaloussi, J. 1996. Developing SGML DTDs. From text to model to markup. Upper Saddle River (NJ), USA: Prentice Hall.

Ranta, J., Statistical metadata - how far (or close) have we got. European Conference on Quality and Methodology in Official Statistics (Q2004). Mainz 24–26 May 2004.

Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Co-dacmos 2004. Project IST-2001-38636.

Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003.