WP. 9
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (ii): Metadata Concepts, Standards, Models and Registries

CONCEPTUAL MODELLING OF STATISTICAL METADATA AND METADATA DATA MODEL IN CoSSI

Supporting Paper

Submitted by Statistics Finland[1]

I. INTRODUCTION

1. Statistics Finland has been starting to implement a statistical metadata concept based on Statistical Information Model called CoSSI (Common Structure of Statistical Information)[2].

2. In modelling of statistical information the methodological starting point of the definition of metadata is that in the conceptualisation of the contentual description of statistical information use is made as far as possible of the concepts characteristic of statistical information, the concepts and concept structures it contains and the logic that allows sufficiently multifaceted and complex concept structures for an exhaustive description of the information content[3].

3. As the used description method of CoSSI allows implementation of complicated structure descriptions, the procedure does not have essentially any factors that would per se somehow force to contract or limit the contentual description.

4. Results obtained when defining statistical information by setting out from the above points of departure are described in the adjacent figure (Figure 1). On the one hand, statistical information has been defined by using a conceptual analysis, the results from which have been depicted as conceptual models of statistical information and, on the other, an analysis has been made of different forms of organising statistical data and

---

[1] Prepared by Heikki Rouhuvirta, Statistical Methodology R&D, Statistics Finland; heikki.rouhuvirta@stat.fi.
[2] Technical description of CoSSI is contained in the definition, see Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.91, Statistics Finland 2003. Also available on the Internet at:
http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf
[3] The foundations and points of departure for the structuring of statistical information, as well as the requirements set on the system for describing it are discussed in more detail in Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

presenting statistical information, which has been used to specify basic models for presenting statistical data. Structural models of data and related data models have been produced for concept models and different forms of organising data, and definitions for these have been implemented in the CoSSI model as multi-level hierarchical (so-called tree-structured) data models[4]. The data models have been documented as XML DTD definitions. The basic method used in the implementation was the ”From model to markup” approach.

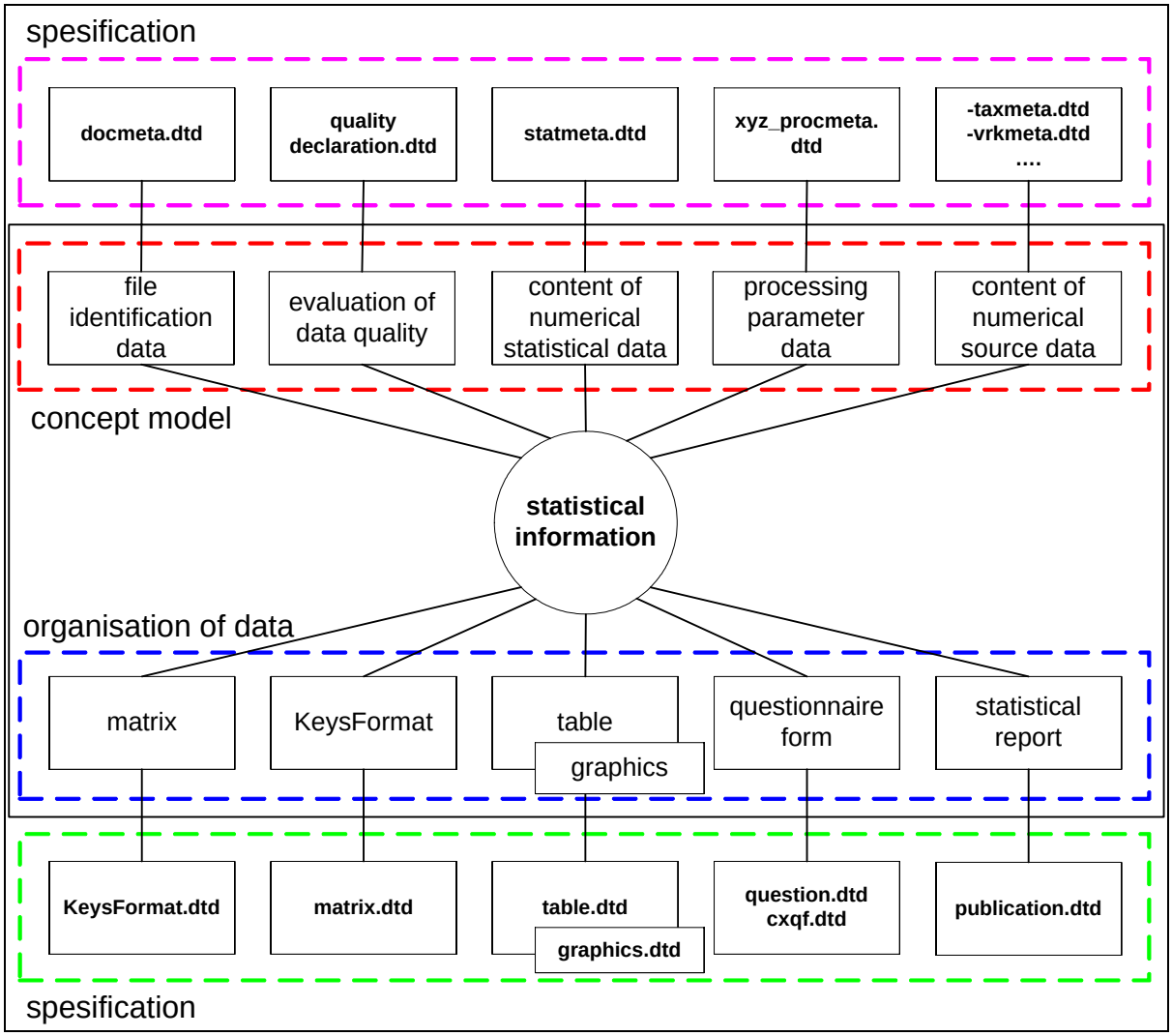

Figure 1. Common Structure of Statistical Information (CoSSI) – parts and entity

5. Basic models for organising statistical data (statistical data files, tables, etc.) are characterised by the fact that the definitions of different forms of organising statistical data allow presentation of the same information content irrespective of the form. Thus, the scope of the data content is not a criterion on the choice of organisation of data but other factors relating to the processing of the data determine the form of organisation that will serve best the production of statistics and the dissemination of statistical information in each case.

6. In CoSSI, all information describing the content, defining, etc., of produced data represent metadata. The following typology of metadata has been used as the metadata frame in CoSSI:

(1) Statistical metadata that are content-specific and necessary for the interpretation of numerical statistical data.

(2) Metadata relating to the identification and archiving of datafiles, which form document metadata.

[4] On demands imposed on the hierarchy of statistical data, see Rouhuvirta, H., Structuring of statistical information and statistical metadata. Codacmos 2004. Project IST-2001-38636.

(3) Metadata concerning processing, of which some belong to statistical metadata as statistical and methodological process data and some belong to the process description as technical metadata required by the used applications.
(4) Technical metadata concerning the process, which contain the technical data required by applications and the metadata used or created in the steering of the project.

7. On the one hand, data obtained from diverse sources for statistical purposes, such as descriptions of data in administrative registers, are based on the own, specific logic of each data source and, on the other, on the availability of data and on the possibility of converting the data into a form where the descriptive information can be electronically attached to the source data and thereby utilised in the production of statistics. Descriptions of source data do not as such form an independent area of their own deviating from statistical metadata, but the descriptive information of the source file is ”included” in one way or another in the statistical metadata as part of the description of the content of the final statistical information.[5]

8. The metadata definitions specifying and describing the contents of statistical information have been technically gathered into the following modules in the CoSSI model:
(1) file metadata (docmeta.dtd)
(2) quality evaluation (qualitydeclaration.dtd)
(3) metadata on statistical information content (statmeta.dtd)
(4) metadata on inquiry (question.dtd)
(5) metadata on register information (e.g. Taxmeta.dtd)
(6) process metadata (e.g. procmeta.dtd).

9. The defining module can be used combined with each other, or as entities supplementing each other dependent of the situation and data description requirements.

II. STATISTICAL METADATA

10. In all situations, the way of processing statistical information is eventually based on the fact that, on the one hand, we have observation units, which in statistics production are also called statistical units. However, on the other hand, besides identification of the observation unit, we also have information produced with different measurement methods on the characteristic of the said unit, which we here refer to as variables for short. This structural characteristic of statistical information (data) is utilised in the CoSSI model to attach and anchor statistical metadata to a variable. Thus, the task of statistical metadata is to describe exhaustively the content and characteristics of the variable for the needs of both producers and users of statistics (see Figure 2).

11. Some of the qualitative information on statistical data describe the characteristics of a variable and some the nature of the entire statistical datafile, and the information is not overlapping or summary in all respects. The metadata relating to a datafile cannot be simplified or assigned to the quality descriptions of the variables it contains but require their own overall quality, i.e. a separate examination of the material entity formed in a certain way. Because of the information relating to quality has been divided into two components and assigned, on the one hand, to the variable insofar as it describes the quality of the variable and, on the other, to the datafile insofar at it describes its characteristics. File-specific quality evaluations are presented in quality descriptions appended to the files.

---

[5] An example of how the descriptive data of an administrative register is handled in the CoSSI framework has been given in Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Final 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 2004. Project IST-2001-38636. Also available on the Internet at: http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf.
An example of how to attach the description of a register into statistical metadata has been given in Rouhuvirta, H., Conceptual Modelling Of Administrative Register Information And Xml - Taxation Metadata As An Example . UNECE Work Session on Statistical Data Editing, Ottawa 2005. Also available on the Internet at: http://www.unece.org/stats/documents/2005/05/sde/wp.3.e.pdf.

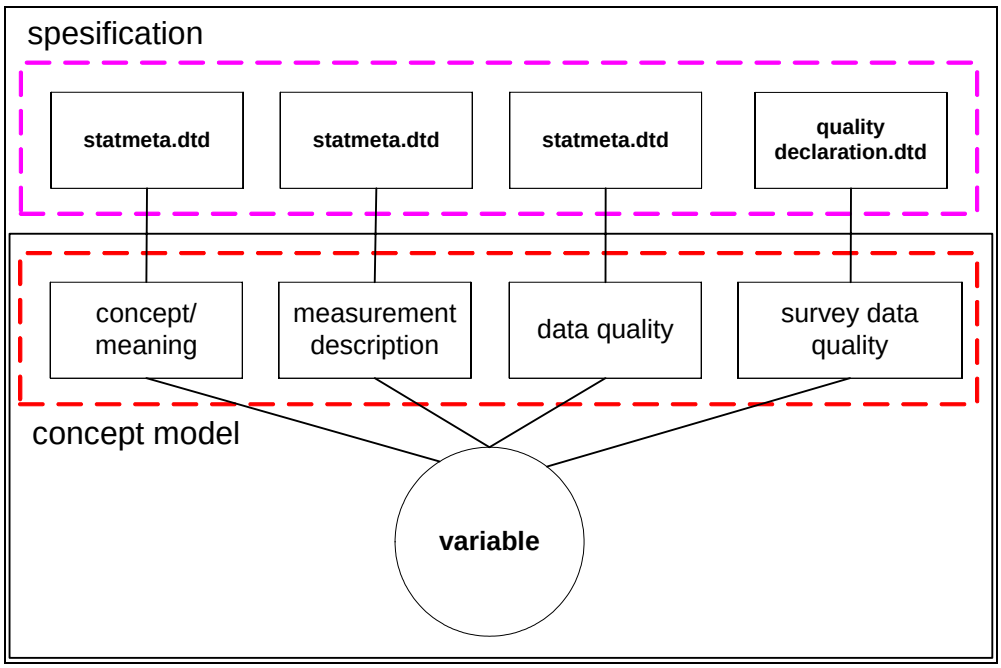

Figure 2. Components of the concepts of statistical metadata and their definitions in CoSSI

12. Variable-centredness also brings the practical benefit that the same metadata description can be used unchanged, and even in the same physical format, in different production stages and in all forms of organising statistical data. This way, many adaptations of the syntax or structure of metadata can be avoided, which might otherwise be necessary for productional reasons.

13. Variable-centredness is a foundation that ensures that metadata are transferred with data to wherever the data are transferred to during statistics production. Irrespective of how the measurement values of measured observation characteristics are handled in different stages of statistics production, the metadata remain the same provided the statistical data themselves are not manipulated in a way that affects their interpretation. Variable-centredness is also a basis whereby descriptions of the contents of administrative and other similar files that are used as sources of statistical data can be combined with the statistical information formed from them in production and in certain cases also with the final description of the statistical information in its dissemination.

14. The description of statistical information at the unit level is comprised of the documentation describing the data of the statistics, which contains statistical metadata by variable and a quality description that contains general methodological descriptions and quality evaluations relating to the data. The variable-specific descriptions of statistical metadata can be supplemented with application-specific process metadata descriptions, in which the technical information required by the application, such as length of data record or its number or character format, can be attached to the metadata descriptions.

III. DATA MODELS OF METADATA

15. The basic conceptual model of statistical metadata is described as a logical data model in the adjacent figure (see Figure 3). The basic, main concepts of statistical metadata relate to the conceptual defining of the content of a variable and to the defining of the measured characteristics. The meaning of a variable is described in a conceptual definition and the matters relating to the measurement in the operational definition of the variable. If the variable is a summary one or one formed in some other manner, the formula used in its formation can be attached to the description of the metadata.

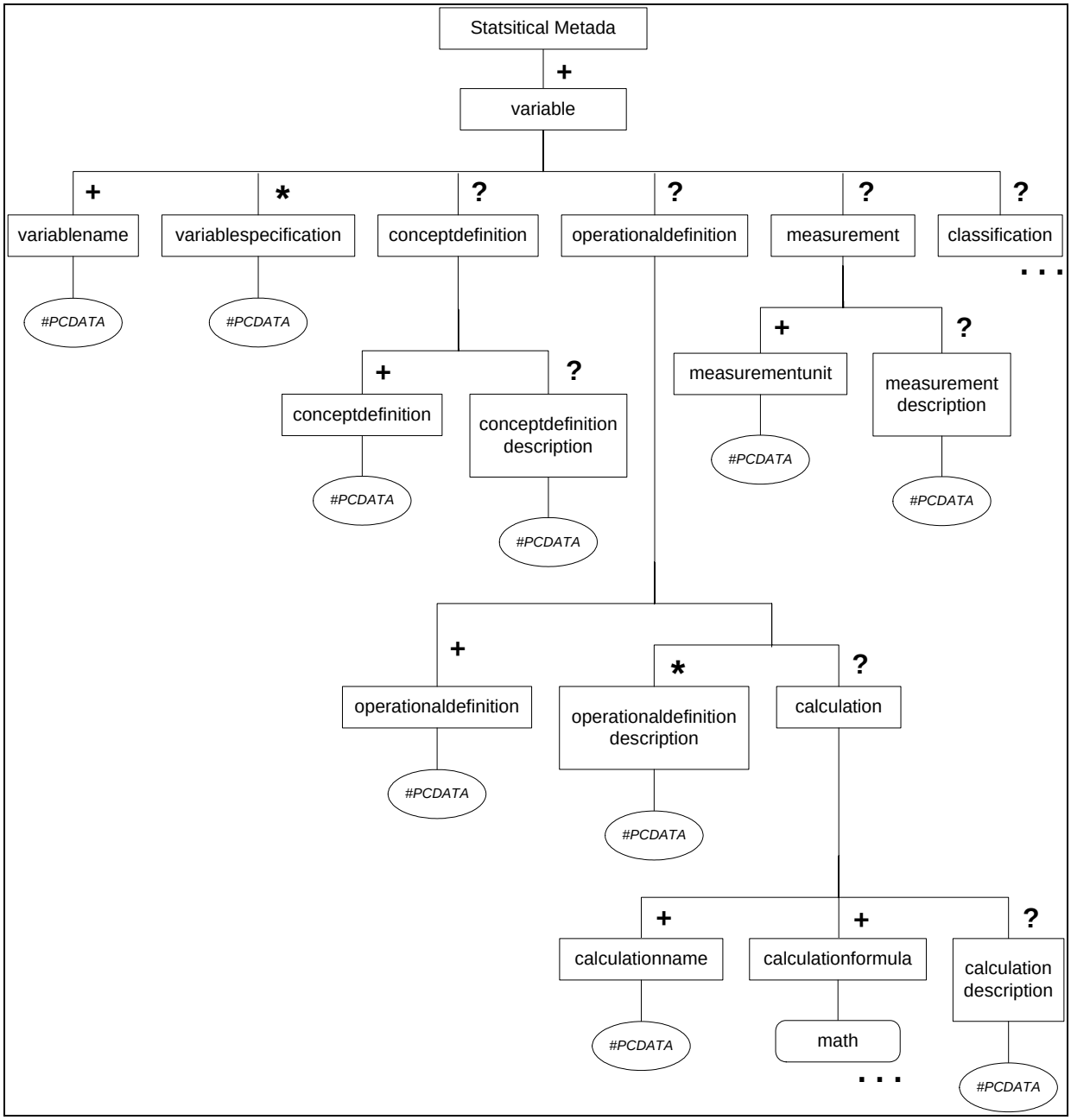

Figure 3. Logical data model of statistical metadata[6]

16. The main elements of statistical metadata descriptions and their purposes of use are presented in Table 1.

17. After initial implementation of the definition of the basic elements of statistical metadata it has become clear that the set of concepts relating to statistical metadata must be enlarged by both conceptualisation of the qualitative data on individual variables and the data that are necessary to steer the processing of statistical data. The steering data assist the communication of the professionals working in production and the realisation of division of responsibilities.

[6] The logical models presented in this context are indicative and detailed, normative data models have been described in the CoSSI definition, see Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.91, Statistics Finland 2003. Also available on the Internet at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf.

Table 1. Basic elements of the concept model of statistical metadata and their purposes of use

Item	Purpose of use
variablename	The name of a variable. Variable name element is used for conceptual naming of variables in natural language. Variable name is not meant to be a code or an abbreviation.
variablespecification	Variable specification is used when the naming of a variable is not enough to describe it. Variable specification element gives a more specific description of the variable.
conceptdefinition	Conceptual definition element contains the conceptual definition of a variable.
conceptdefinition description	Conceptual definition description is used when the information in the conceptual definition element is not enough to clarify the conceptual aspects of a variable.
operationaldefinition	Operational definition element contains a written operational definition of a variable.
operdefinition description	Description element of an operational definition includes a written description of the operational definition. The description is given in natural language. Description of an operational definition is used when the information in an operational definition element is not enough to clarify the operational aspects of a variable.
calculationname	Name of a calculation. If a calculation is given it must be named. It is possible to give the name of the used method without giving the actual calculation formula. The name can be a generic or a case-specified name of the method.
calculationformula	The actual calculation formula is given here in MathML format.
calculation description	Describes a calculation formula. Calculation description is used when the information in the calculation name and calculation formula is not enough to clarify the composing of a variable.
measurementunit	This element names the measurement unit of a variable. The measurement unit is given as a standardised Finnish abbreviation (at Statistics Finland).
measurement description	The description is used to clarify the measurement if the measurement unit is not clear enough.

18. Extensions made at the first stage to the logical data model of statistical data to serve the said purposes are presented in the adjacent figure (see Figure 4)[7].

19. The new elements attached to a variable for production purposes are:

-field name

Technical identifier: Technical identifier enables the use of short names of variables in different information technology environments where the use of long, natural names is not always possible because of technical reasons. In addition to, and separate from this, a variable has a universal identifier (ID).

-survey id

Survey identifier: Statistics departments collect data for more than one statistical survey simultaneously and/or form different ”versions” of a variable. To help identification in production, a variable can be given a survey identifier in the form of a set of characters.

-variable code

Variable code: Variable code facilitates denoting hierarchical variables so that the code can be used in data processing and output. An example would be the long list of income categories in income distribution statistics, in which income variables have been given numerical codes with which different income concepts can be summed up semi-automatically. In this respect, the numerical codes of the income categories could be compared to the sets of codes used in regional classifications. The numerical codes of the variables are included as separate elements in the model because their purpose of use is different from that of the technical or ID codes.

-derive rules

Derive rule: Derive rule is a productional element into which the compilers of statistics can record in their own way in statistical jargon the rule by which a variable is formed.

[7] The extensions shown in this context are due to be implemented into version 2.0 of the CoSSI definition during 2006.

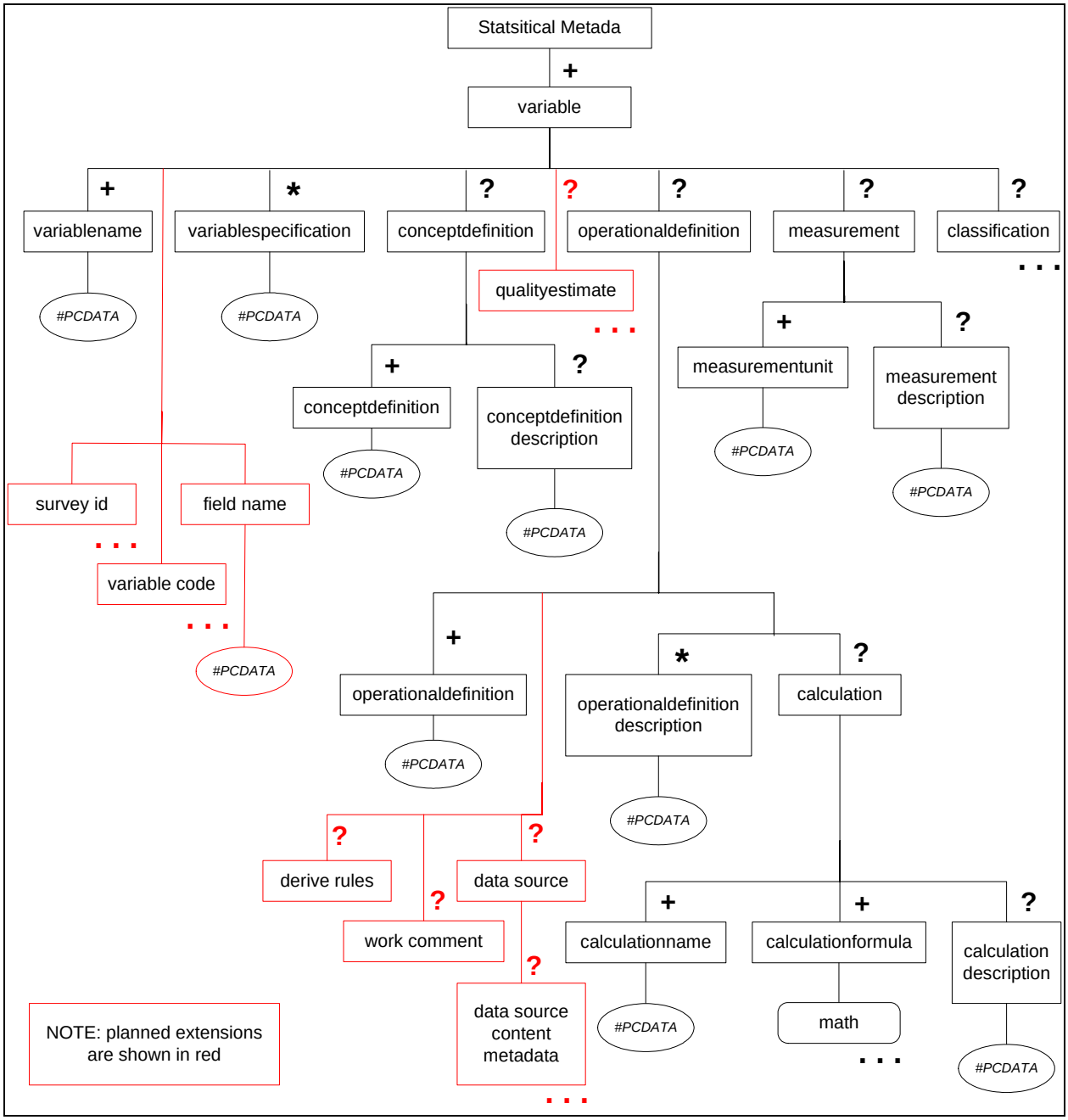

Figure 4. Extensions of the logical data model of statistical metadata

The intention is that these preliminary expressions will be used to develop an operative definition of a variable, which can be registered as an operational definition of the variable in terms of its accuracy and understandability while at the same time retaining the original derive rule expressed in statistical jargon for productional purposes. The derive rule functions at the same time as a common definition document for the compiler of statistics and for the application developer/programmer.

-work comment

Work comment: Work comment is intended to be used in the supervision of the work of compilers of statistics and as a production check list.

-data source

Source of data on variable: A link or direct reference can be given to a variable to either an external data source, such as an administrative register, or to a question in a question database of data collected in-house.

-metadata on the content of source data

Description of the content of source data: Description of the content of an external data source can be attached here, if the structural description of the data is known, as is the case with taxation data if they have been described according to taxmeta.dtd or if the description of a question relating to survey data collected in-house is in the format specified in question.dtd. A description complying with question.dtd contains the original question text and the values and descriptions of the reply alternatives to it.

20. The above-described data are primarily meant for production purposes, and it is not the intention to include them in systems for disseminating statistical information. Descriptive information on data sources, which in itself is public information but whose inclusion in the dissemination of statistical information is subject to a separate agreement with the original data producer is, of course, a borderline case. This procedure must be followed irrespective of the fact that the data themselves can be used for statistical purposes. In an ideal case, description of the input data can be included as part of statistical metadata when understandably expressed.

21. Once the data are completed the purpose of the data evaluating the quality a variable is to help the users of statistical information to assess and use the statistics correctly. A preliminary data model for data evaluating variable quality is presented in the adjacent figure (see Figure 5).

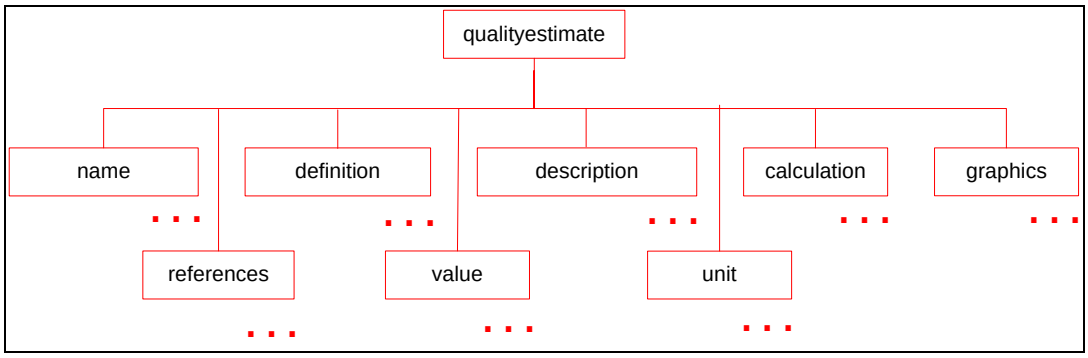

Figure 5. Extension of the logical model of statistical metadata – quality evaluation of variable

22. The quality evaluation data attached to a variable comprise the following elements:

-quality estimate/quality index/indicator [quality judgement],
-name of parameter,
-method or formula for calculating parameter,
-definition of description of parameter,
-reference to possible methodological source,
-description or interpreting instructions,
-calculation result/value/result value,
-calculation unit of parameter and
-graphic depiction or presentation of result or result value.

23. The central component of statistical metadata is description of the classification of the variables. In the concept model, matters relating to used classifications have been described in two ways. On the one hand, the used classification standard can be identified or, alternatively, the used category values and their importance can also be described. The entire element concerning classification in the concept model is presented in the adjacent figure (see Figure 6).

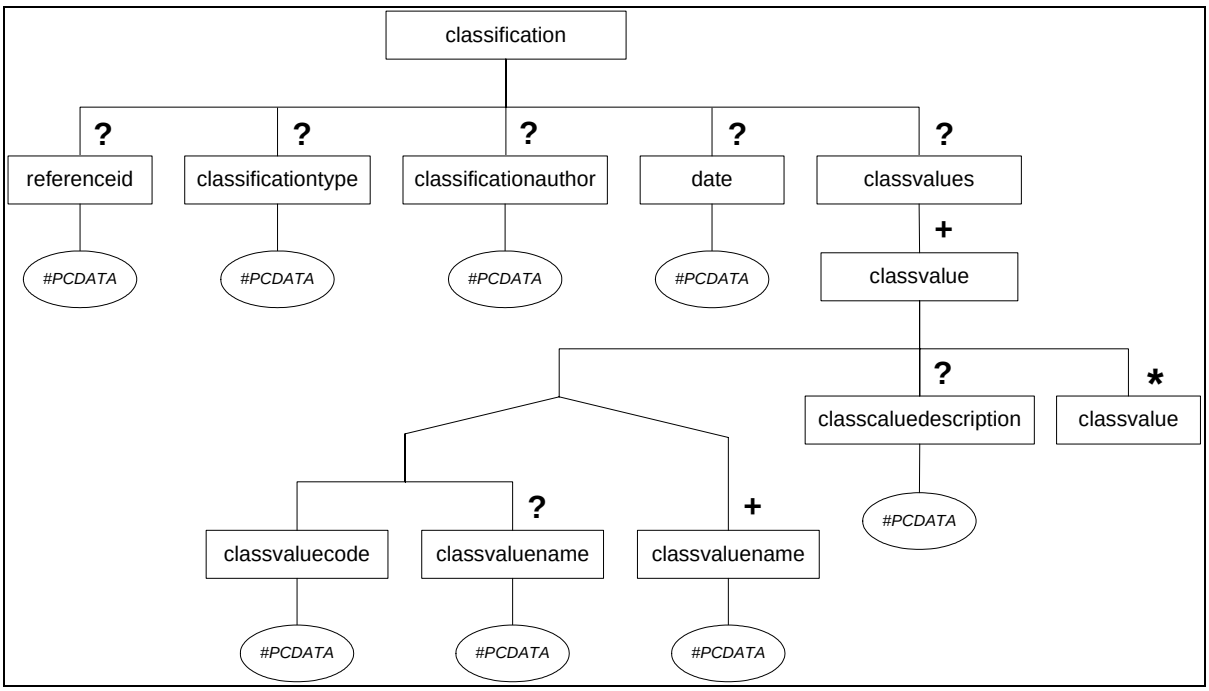

Figure 6. Logical data model of statistical metadata - classifications

24. The purposes of use for the metadata elements describing statistical classification data are presented in Table 2.

Table 2. Basic elements concerning statistical classifications in the concept model of statistical metadata and their purposes of use

Item	Purpose of use
classification	The classification element of statistical metadata can be used in two ways. First, when a standard classification is applied the element can be used to give the identification data (codes, names, etc.) of the classification. Second, the element can be used to describe an own classification for which general identifiers need not be given
referenceid	Reference identifier of the used standard classification. Classification identifier can be used as a direct link to a classification.
classificationtype	Classification type is used for descriptive purposes of standard classifications.
classificationauthor	Author of the classification.
date	Date when classification was published.
classvalue	Collecting element for all information of the class value. Hierarchy level of the class value is described with a hierarchy level attribute. Description of this element's content model: The obligatory information that must be given on the class value is either its code or its name. Both can also be used simultaneously.A separate description can be attached to the class value. The class value may also contain subclasses that may contain further subclasses. At the moment the hierarchy is limited to six levels
hierarchylevel	Attribute describes at which level in the classification hierarchy this item is.
classvaluecode	Code of a class value
classvaluename	Element contains the name information of a class values in textual format.
classvalue description	Class value description is used when the information in a class value name element is not enough to clarify the conceptual issue of the class value.

25. After the first implementation of the definitions of the basic concepts of classification data listed above it has become quite clear that the classification description must be extended to allow both work comments recorded for production purposes and descriptions of the formal data related to the formation of classifications. The planned additions are presented in Figure 7.

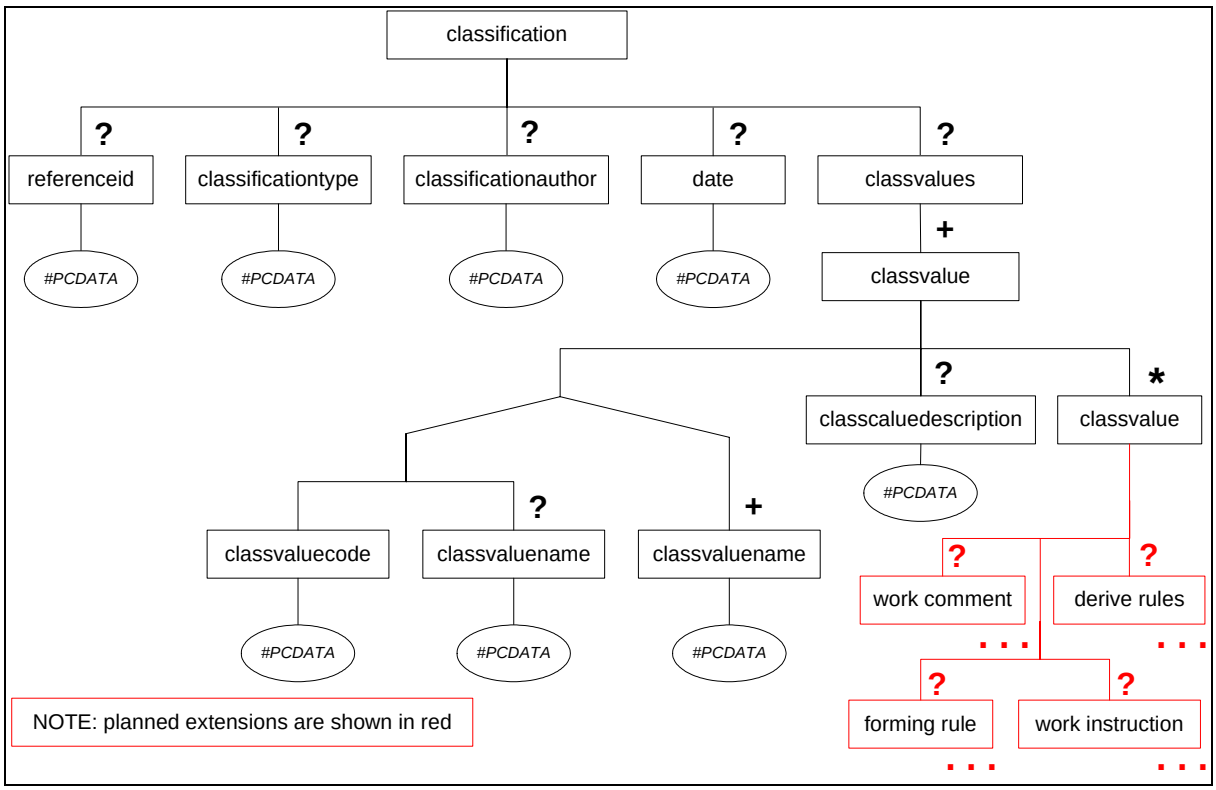

Figure 7. Extensions concerning classification data in the logical model of statistical metadata

26. The elements to be added to classification descriptions are:

-derive rules

Category value formation rule: The formation rule can contain as data the formation rule expressed as a formula together with its description.

-forming rule

Inference rule: Inference rule is a formation rule formulated for production purposes, which is often expressed in statistical jargon, but forms a basic common document for the compiler of statistics and for the application developer/programmer. Final category description is formed from the inference rule, but inclusion of description of the inference rule as a separate element has been retained because this way it can also be documented and stored separately.

-work comment

Work comment: Work comments are intended to be used in supervising the work of statistics compilers and as a check list in production.

-work instruction

Work instruction: Work instructions are intended for the documentation and storing of formal algorithms (at Statistics Finland these are referred to as ADP rules on the forming of a category value or similar).

27. The quality description of a datafile follows the approved manner for producing quality descriptions at Statistics Finland. The concept model of the planned quality description is presented in Figure 8.

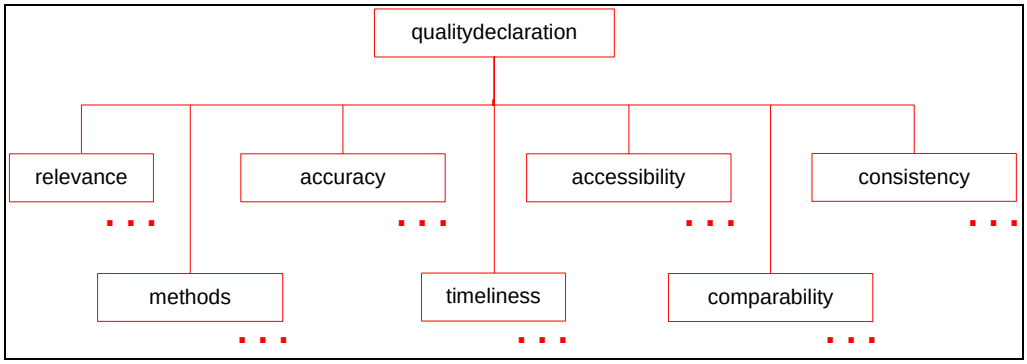

Figure 8. Logical data model of quality description

28. Purposes of use for the central concept elements of a quality description are presented in Table 3.

Table 3. Basic elements of the concept model of a quality description and their purposes of use

	Item	Purpose of use
1.	relevance	Relevance of statistical data
1.1.		Produce a detailed summary of the product's information content and end use. Identify the phenomenon that this set of statistics is designed to describe and explain its history.
1.2.		Introduce concepts that are important to understanding the statistics, classifications used or object of study, to identifying the data collector and informants.
1.3.		Describe the any acts, decrees and recommendations upon which the statistics are based.
1.4.		Assess the relevance of the statistical information produced in relation to customer needs, and how any changes in the phenomenon concerned have been taken into account in compiling the statistics.
2.	methods	Description of the methods used in statistical surveys
2.1.		Describe methods precisely, e.g.the methods applied, i.e. the population of the statistics, the materials used, the survey design (census survey or sample survey), the sampling design, data collection method, editing, imputation, the use of weighting coefficients in sample surveys and estimation methods required by the final results
2.2.		Justify the methods used and any changes made (including an assessment of the impacts of those changes upon time series).
2.3.		Methods descriptions identify the data sources used in statistics production (also for auxiliary information).
2.4.		Review the whole process of statistical survey.
3.	accuracy	Accuracy of information
3.1.		Demonstrate that the statistics measure what they are supposed to measure.
3.2.		Report on all facts that may have a bearing on the reliability of the statistics. Also mention key uncertainty factors, i.e. possible sampling and non-sampling errors.
.3.3.		Estimate the correspondence between the target population and the population of interest and the quality of the sampling frame used.
3.4.		Describe the main uncertainty factors, i.e. possible sources of error, and assess their impacts on the estimates published: – Sampling errors, – Non-sampling errors: – Coverage error, – Measurement error, – Processing error, – Non-response error.
3.5.		Using the main classifications employed in the statistics, tabulate statistical parameters for the estimates, such as standard deviations that take the sampling design into account, mean square errors (MSE) and parameters

		estimating the efficiency of the sampling design (deff)
3.6.		Interpret tables produced in 3.5.
4.	timeliness	Timeliness and promptness of the information published
4.1.		Indicate the point of time or period that the statistics describe.
4.2.		Indicate whether the information is preliminary or final.
4.3.		Where necessary examine how time series data have changed over time (e.g. on account of seasonal adjustment).
5.	accessibility	Accessibility and clarity of information
5.1.		For statistics where the data constitute comparable time series, indicate the length of the time series available.
6.	comparability	Comparability of statistics
6.1.		Describe the comparability of the statistics over time and with other materials.
6.2.		Examine changes that have affected comparability and their significance, e. g. in the statistics production process, survey design concepts and classifications.
7.	consistency	Consistency
7.1.		Assess the consistency of the statistics in comparison with other statistics on the same subject. For example, examine differences in their concepts and data collection processes and assess their impacts.

29. From the point of dissemination of statistical information and exploitation of statistical datafiles, essential metadata are also contained in the metadata concerning individual files. The metadata defining a file contain information on its producer, subject, identifiers, etc. (see Figure 9.).

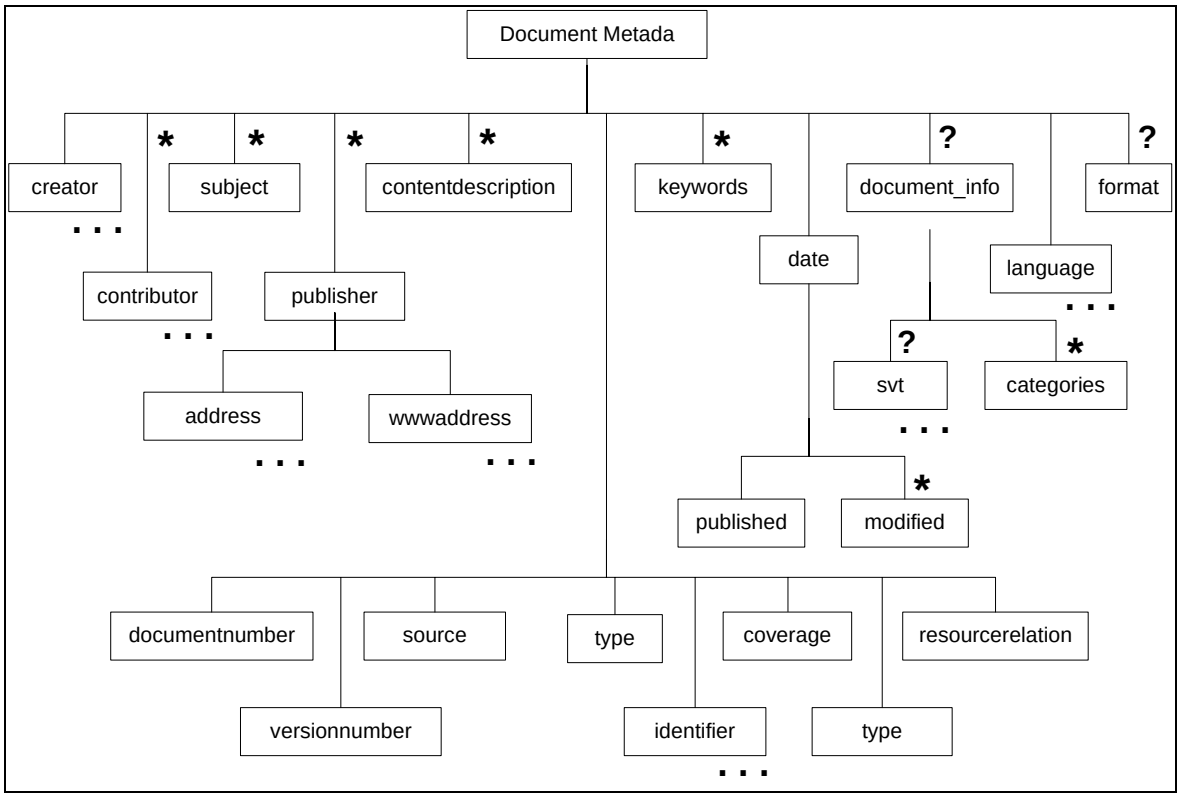

Figure 9. Logical data model of document metadata

30. Purposes of use for the central concept elements of the document metadata are presented in Table 4.

Table 4. Basic elements of the concept model of document metadata and their purposes of use

creator	Equivalent to Dublin Core element Creator: An entity primarily responsible for making the content of the resource. Examples of Creator include a person, an organisation or a service. Typically, the name of a Creator should be used to indicate the entity. At Statistics Finland Creator is a logical element and the name of the creator is given in the child element person. In statistical publications the creator is the person who can give more information about the concerned topic. In other publications (i.e. research reports) the creator is the person responsible for the text. There can be only one creator for any one document. If other persons have contributed to the making of the document, they are mentioned in the contributor element. When information about the creator is copied to the bibliographical information system using Dublin Core, and the document constitutes Official Statistics, then Statistics Finland uses its organisational name as the creator.
subject	Equivalent to Dublin Core element Title: A name given to the resource. Typically, Title will be a name by which the resource is formally known. At Statistics Finland Subject is the title text of a document. It means that the subject element in this module repeats the title element of the parent DTD (e.g. publication.dtd or table.dtd). For example, if a document is a table (table.dtd), then subject is the content of the title element of the table document. When subject information is copied to the bibliographical information system using Dublin Core, then the subject is the same as the title.
keywords	Equivalent to Dublin Core element Subject and keywords: Topic of the content of the resource. Typically, Subject will be expressed as keywords, key phrases or classification codes that describe the topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme. At Statistics Finland the first keyword in an Official Statistics publication is the category of the document, so it is repeated here. Otherwise (e.g. in research reports), keywords can describe the publication's frame of reference.
contentdescription	Equivalent to Dublin Core element Description: An account of the content of the resource. Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content. Not currently used by Statistics Finland's system architecture. This is an optional element for possible future needs.
publisher	Equivalent to Dublin Core element Publisher: An entity responsible for making the resource available. Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.Use at Statistics FinlandPublisher is a logical element for connecting Statistics Finland's organisational information to the Dublin Core publisher information. The name of the publisher, in different languages, is always Statistics Finland. With statistics produced by other statistical organisations, the publisher is the organisation concerned.
contributor	Equivalent to Dublin Core element Contributor: An entity responsible for making contributions to the content of the resource. Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.Use at Statistics FinlandContributor is a logical element and the information about a contributor is given in the child element person. Other persons besides the creator who have contributed to the making of the document. In Official Statistics this means the secondary source of additional information.
date	Equivalent to Dublin Core element Date: A date of an event in the lifecycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601.At Statistics Finland Date is written in the form DD.MM.YYYY according to the Finnish standard.
type	Equivalent to Dublin Core element Type: The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary). To describe the physical or digital manifestation of the resource, use the FORMAT element. At Statistics Finland Types of document are: table, publication (Official Statistics), statistical news, figure, article, research report, handbook or statistical metadata document. If the type of document is recearch report, then the identification number of the document is given in the element documentnumber.
format	Equivalent to Dublin Core element Format: The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types defining computer media formats). In Statistics Finland's systems the format is either XML or XHTML. Format is currently only used in electronic publications.
language	Equivalent to Dublin Core element Language: A language of the intellectual content of the resource. Recommended best practice is to use RFC 3066 which, in conjunction with ISO639, defines two- and three primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian", and "en-GB" for English used in the United Kingdom. At Statistics Finland this element is a logical element for the main language and other language elements.

document_info	Document_info is a logical element that contains Statistics Finland-specific metadata information. Series and category of Official Statistics are covered here.
identifier	Equivalent to Dublin Core element Identifier: An unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN). This is a logical element for the different identifiers of a document. (URN | URL | ISBN | ISSN | DOI | documentnumber)
URN	Uniform Resource Names. Currently not in use at Statistics Finland.
URL	Uniform Resource Locator. Fixed URLs facilitating source references should be used at Statistics Finland.
ISBN	International Standard Book Numbers. Statistics Finland uses its own ISBN series.
ISSN	International Standard Serial Number. Statistics Finland uses its own ISSN series.
DOI	Digital Object Identifier. Currently not in use at Statistics Finland.
documentnumber	Number issued for a publication to identify it in a certain series. If a publication B32constitutes Official Statistics, then it uses its own series number, which is described in the element number.Statistical news have an identifier in the form of "year:number" (e.g. 2002:157) and research reports belonging to a given series have their specific number sequence.
rights	Equivalent to Dublin Core element Rights:Information about rights held in and over the resource. Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.Use at Statistics FinlandThe name of the organisation that has the rights to the content of the document.
coverage	Equivalent to Dublin Core element Coverage:The extent or scope of the content of the resource. Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges. Currently not in use at Statistics Finland.
resourcerelation	Equivalent to Dublin Core element Relation:A reference to a related resource. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system. Currently not in use at Statistics Finland.
source	Equivalent to Dublin Core element Source:A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.Use at Statistics FinlandSource is reserved for tables. It identifies the statistical data matrix from which the table is produced. Not in use for publications.
person	Person is a logical element that includes a personal information set describing the creator of and contributors to the document.
organisation	Organisation is a logical element containing an information set in one language about the organisation that has produced the document.
published	Logical element for the official publishing date and time of a document. The date contained by the element will not be changed even if the document is edited further or errors in it are corrected and the document is republished. In such cases the modified date is changed.
modified	Logical element for the dates and times of modifications made to a document. When a document is edited further or corrected for errors and the document is republished, the dates of republishing are recorded in the modified element. The chronologically latest element gives the date of the most recent amendment or update.
day	Date is written in the form DD.MM.YYYY according to the Finnish standard.timeTime is written in the form HH:MM:SS
main_language	Defines the main language of a document, which is also the default language of an electronic document. Main language can be any language. The practice is to use RFC 3066 which, in conjunction with ISO639, defines two and three primary language tags with optional subtags.fi, en, se
other_language	Defines one or more secondary languages of a document. The practice is to use RFC 3066 which, in conjunction with ISO639, defines two and three primary language tags with optional subtags.fi, en, se
svt	Svt is a logical element containing serial information on Official Statistics publications (SVT). SVT is used to identify publications published in the Official Statistics of Finland (OSF) series (name and number of the series, year, version number) and to identify Official Statistics tables.

categories	Categories element describes the topic of the statistical information. The attribute gives the topic (category) as a three-letter abbreviation. The whole name of the category and its different language versions are given as entity files.
type	Type is the three-letter abbreviation describing the category of the statistical information.
first_name	Person's first name.
surname	Person's surname.
position	Person's official title at Statistics Finland.
email	Person's (or organisation's) e-mail address. Use at Statistics FinlandE-mail address is given in Finnish language in the form "firstname.lastname@tilastokeskus.fi" and in foreign languages "firstname.lastname@stat.fi". Language attribute (xml:lang) must be given according to the document languages.
phonenumber	Person's (or organisation's) phone number.
fax	Person's (or organisations) fax number.
orgname	Name of the official statistical organisation in one language. Language is defined in the parent element organisation. Tilastokeskus, Statistics Finland or Statistikcentralen
division	Name of the official statistical organisation's sub-unit (publishing sub-unit) in one language. Usually means the responsible unit that has published the statistics. Language is defined in the parent element organisation.
address	Street address of the organisation.
wwwaddress	One www-address of the organisation in one language.
seriename	The identifier part is comprised of three elements: series name, year and number. Seriename identifier is used to identify the tables, figures and publications that come within the scope of Official Statistics of Finland. At Statistics Finland, the series division of Official Statistics currently differs from statistical topics. The seriename element depicts the topic of a publication in Official Statistics serie. The possible category values are: Housing; Living conditions; Energy; Prices and costs; Government finance; National accounts; Trade; Education; Culture and the media; Transport and tourism; Agriculture, forestry and fishing; Justice; Wages and salaries; Financing; Construction; Manufacturing; Health; Science, technology and research; Income and consumption; Labour market; Foreign trade; Elections; Population; Population census; Environment and natural resources; Enterprises.
year	Year of publishing of Official Statistical information.
number	Annual sequential number within a series by document type. Formed series-specifically for all publications, tables, etc., of the series concerned. Tarkoitus käyttää ensisijaisesti taulukoiden yhteydessä. Sisältö joudutaan mahdollisesti määrittelemään uudelleen, kun painetut julkaisut häviävät.
versionnumber	This element is the version number of a document. Tarkoitettu käytettäväksi lähinnä taulukoiden ja kuvioiden kanssa.

IV. SOME CONCLUSIONS

31. According to our initial experiences, application of a structured information model to the conceptualisation of statistical metadata has opened new possibilities for exploiting the new technologies developed for easy handling of information in text format. XML can be regarded as representing such technologies. More efficient processing of text-format information facilitates management of richer statistical metadata. This aspect can be exploited equally well in the production of statistics and in the dissemination of statistical information.

32. The practical benefits brought by the application of structuring to the management of metadata include the ease with which the data model can be expanded and the extensions can be technically implemented. The editability can be utilised to develop as consistent and all-embracing content specification as possible for statistical metadata.

33. Besides the scope of its content and its flexibility, as a frame of reference for metadata the CoSSI model examined here differs quite essentially from the metadata systems conventionally used at Statistics Finland in that

-it makes it possible to change from decentralised to centralised management of metadata in which the producer of statistics can control the correctness of the metadata concerning the data material, and which can also be used in the dissemination of statistical information, and
-when receiving numerical statistical data, in the same connection the users also receive the metadata that are essential in their interpretation, and instead of untargeted metadata in separate reference volumes or other similar sources, metadata can be presented immediately adjacent to numerical statistical data.

34. Examined from the perspective of production, the point of departure in the modelling and organisation of metadata could have been its attachment to the numerical value of data. However, the now implemented attachment of metadata to the variable instead of the data value facilitated the use of a simpler and more informative data model relative to the scope of the data content, and simplified and rationalised the management of data. Attachment of statistical metadata to the structure of statistical data through a variable is technically considerably simpler to implement than linking of metadata to individual data values in an information system, whatever the available production technology.

35. In practice, the structured model of statistical information represents a real alternative as a frame of reference for statistical metadata, both in respect of its approach and concept defining, and at the moment there appears to be no specific need to change the developed basic solution. Indeed, the needs for further development concern primarily extensions of the data content along the lines described above. Moreover, we are endeavouring to improve the functionality of the technical solutions of statistics production so as to make the use of statistical metadata effortless and easy in different stages of production. These kinds of solutions serving the production of statistics include creation of user interfaces with statistical metadata into such tabulation applications as SAS and SuperStar.

References

Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at:
http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

Rouhuvirta, H., Conceptual Modelling Of Administrative Register Information And Xml - Taxation Metadata As An Example . UNECE Work Session on Statistical Data Editing, Ottawa 2005. Also available on the Internet at: http://www.unece.org/stats/documents/2005/05/sde/wp.3.e.pdf.

Rouhuvirta, H., Structuring of statistical information and statistical metadata. Codacmos 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. Available on the web at:
http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf

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. Codacmos 2004. Project IST-2001-38636. Available on the web at:
http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf