WP. 20
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 (iii): Metadata and the Statistical Cycle

DISSEMINATION OF STATISTICAL DATA AND METADATA - PROCESS BASED ON COMMON STRUCTURE OF STATISTICAL INFORMATION (COSSI)

Supporting Paper

Submitted by Statistics Finland[1]

I. INTRODUCTION

1. Statistics Finland has been developing XML-based data dissemination for a couple of years now. The dissemination system is based on the model of common structure of statistical information (CoSSI), and the dissemination is based on XML documents compatible with it. The CoSSI model covers different ways of statistical data organisation (statistical data matrix and statistical table), statistical publications (monthly and quarterly publications, press releases, etc.) and quality declarations. The structuring of the metadata connected to statistical data is also implemented within this system.

2. The metadata part in the CoSSI model is divided into document metadata, statistical metadata and processing metadata. Document metadata is information about the producer of the document, the document's content, date, statistical topic, etc. Statistical metadata is information vital for the interpretation of numerical statistical information, and describe the variables in a statistical table or matrix. This metadata information is useful for the user in the dissemination process by helping the interpretation of statistical figures, and for the producers of statistics when metadata are transferred between statistical production stages. It could be also used for bilateral exchange of statistical information between statistical agencies.

3. The metadata, publications, tables and matrixes will be stored in an XML database. This database is also an archive for all published information. Statistical publications are prepared in the statistics departments of Statistics Finland. People responsible for publications will have access to the XML database via an XML editor. Using this XML editor they can make publications, and include tables, figures and metadata in them, and also update and write new metadata documents describing their statistics. So a publication in XML format, defined by the CoSSI model, includes all the statistical data, metadata, text, figures and language versions in a single XML file. This XML file will then be automatically converted to HTML files for the website, PDF files for printing or to some other appropriate format.

[1] Prepared by Harri Lehtinen (harri.lehtinen@stat.fi).

4. PC-Axis is also starting to use the CoSSI model as XML format. For this the CoSSI model contains a special processing metadata section called pxmeta. This PC-Axis metadata information can also be stored in XML format in the same XML database as all the other information. By doing this, database (PX-Web) publishing can also be incorporated into the same process as publication production.

5. Using XML and XML tools, such as XML database and XML editor, we can unify the production of information for different publication channels. This also gives us the opportunity to include rich metadata in statistical information and publish metadata along with the statistics they describe.

II. COSSI - COMMON STRUCTURE OF STATISTICAL INFORMATION

6. At Statistics Finland we have been developing a common structural definition of statistical information (CoSSI [2]), which covers different ways of statistical data organisation (statistical data matrix and statistical table), publications, and within which the structuring of the metadata connected to statistical data is also implemented.

7. The CoSSI defines the structures of statistical data (matrices and tables), metadata (document and statistical metadata, and quality declarations), and publications. XML DTDs have been selected as the technical means for implementing these structures. The CoSSI model is comprised of several DTDs that can be modularly combined for different types of documents. The basic document types are a statistical table (CALS), a statistical matrix (XDF) and a publication. These documents are XML documents that are compatible with the CoSSI model and also contain the metadata and the language versions necessary for describing a set of statistics.

A. Metadata in CoSSI model

8. In the CoSSI model, metadata are divided into four categories:

a. Document metadata
b. Statistical metadata
c. Quality declarations
d. Processing metadata

9. The division is based on the content and character of metadata. Document metadata describe the content of a document, its creator and identifiers connected with the document. Statistical metadata contain descriptions of the variables that are present in statistical data and tables, calculation rules and any classifications that may apply to a variable. The quality declaration is a standard format description of the data collection, the data and the applied statistical methods. Processing metadata are metadata for statistical software applications.

B. Document metadata

10. The document metadata module has all the elements of DublinCore, as well as some metadata specific to Statistics Finland. The document metadata DTD cover the metadata that are needed for the publications, tables and matrices of Statistics Finland (electronic and paper dissemination).

11. Document metadata provide information about the person and organisation having produced the document, the content of the document (e.g. subject, keywords, language, source) and information specific to Statistics Finland, such as the series and category of Official Statistics of Finland. Productional metadata, processing instructions, etc., are not contained in document metadata.

---

[2] 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.

12. Document metadata is a module that is used in tables, matrices, statistical metadata, and publications for describing document metadata. The document metadata module can also be used for the production of information for bibliographical information systems.

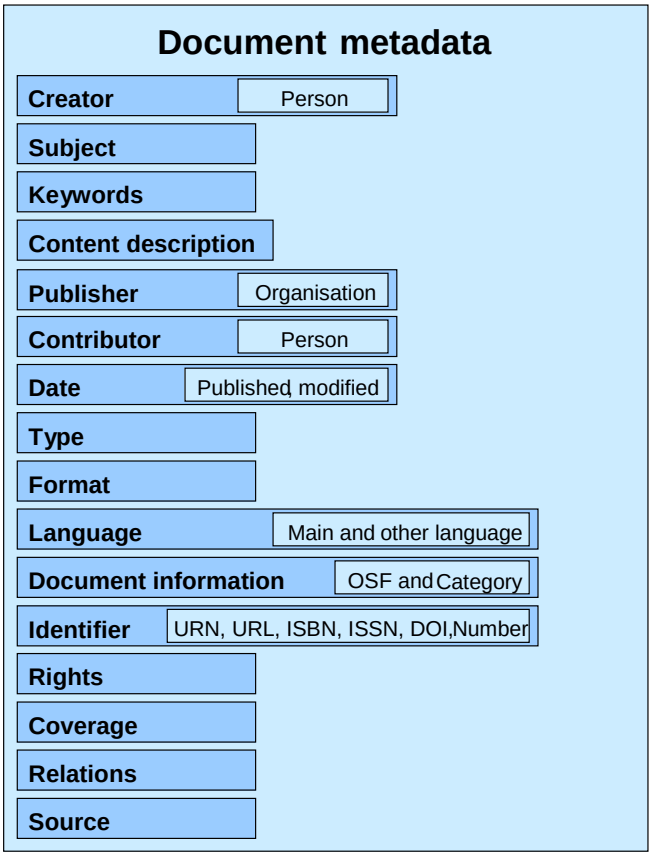

Figure 1. Content model of document metadata

C. Statistical metadata

13. Statistical metadata are data (matrix or table) specific. A statistical metadata document describes the variables, and their operational definitions and classifications in matrices and tables. If data (matrix or table) are changed as a result of tabulation or other procedure, these changes are added to the statistical metadata document. Thus, if a variable is edited, the pertinent metadata are added to the statistical metadata.

14. 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 model this has been solved as part of the structuring of statistical information.

15. 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 are described in another component of the CoSSI model that covers metadata concerning documents and file copies.

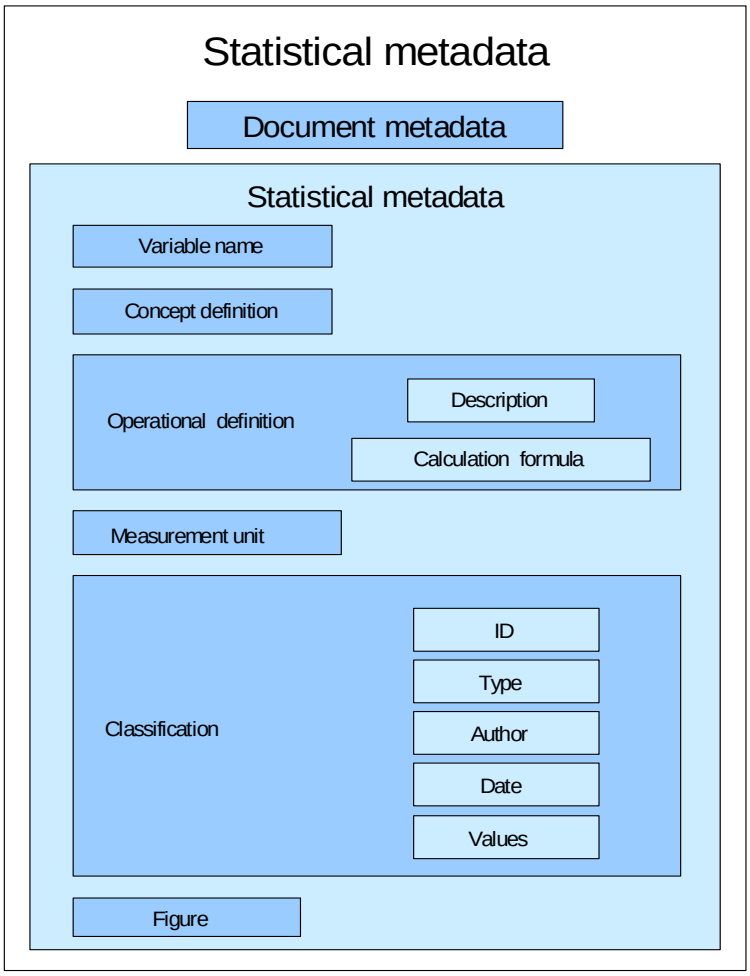

Figure 2. Content model of statistical metadata

16. The model of statistical metadata is very simple but rich in information content. Its main benefit to the production of statistics is that statistical metadata retain a consistent structure right from a survey questionnaire to the eventually disseminated statistics. Because of their simple structure, statistical metadata can be easily exploited during different stages of statistics production and the same model of statistical metadata can be used in all stages from the collection of data to their publication and dissemination. As the same model is used, it does not impose any necessity to re-enter, convert or otherwise re-process statistical metadata, which remain unchanged and go together with the numeric data they describe from the start to the finish of the data processing procedure.

17. The second major benefit is the extendibility of the data, for instance, in situations where statistical standard classifications are used. The model of statistical metadata provides a set and standardised ways of making standard classifications part of them.

D. Quality declaration

18. Quality declaration, or quality metadata, describe the following aspects of statistics:

a. Relevance of statistical data
b. Description of the methods used in statistical surveys
c. Accuracy of information
d. Timeliness and promptness of the information published
e. Accessibility and clarity of information
f. Comparability of statistics
g. Consistency

19. In publications, tables and matrices, quality metadata can be provided as a quality declaration, which is a module in the CoSSI model.

E. Processing metadata

20. Processing metadata are metadata intended to be used by a certain software application to guide its functioning. Examples of these would be certain PC-Axis keywords that instruct the application to show the figures in a table at the accuracy of certain number of decimal digits or create the heading of a table automatically according to predefined rules. The question so far studied in respect of processing metadata is how data in the PC-Axis file format could be included in an XML format document compatible with the CoSSI model and which PC-Axis keywords should then become processing metadata. In future, it would be possible in principle to expand the CoSSI model with the processing metadata of other statistical applications, such as SAS and SuperStar, depending on the production environment.

III. XML-BASED PUBLISHING SYSTEM

21. Statistics Finland has been developing XML-based publishing system which sets out XML documents, publications, tables, matrices and metadata that are compatible with the CoSSI model. The system converts these XML documents automatically into the formats required by different dissemination channels.

22. An output format compatible with the tables and matrices of the CoSSI model is needed for statistical software applications for XML-based publishing. The main applications Statistics Finland uses are SAS and SuperStar, and PX-Edit for PC-Axis tables. At the moment PC-Axis tables in matrix and table formats can be produced with PX-Edit and output format for matrices and tables has also been developed for SAS. In the newest version of SuperStar there will be an output in the CoSSI table format (CALS).

23. The actual production of publications takes place at the statistical operating units where statistical experts write the text, select the tables for the publication and produce the statistical graphics. Epic software was selected as the editor for XML-based publishing, and was tailored during year 2005 to function as the production editor for publication documents conforming with the CoSSI model. In the tailoring the user interface of the editor was made as user-friendly as possible and functions were added to it that support e.g. importing of external XML tables compliant with the CoSSI model into a publication, production of language versions, completion of metadata and writing of text. Technically, XML is hidden in the editor, so the editing environment is quite similar to that of familiar word processing programs.

24. An XML database into which statistical metadata compatible with the CoSSI model are saved as XML documents has now been taken into test and piloting use. As tables are made, descriptions of the variables selected for them can be retrieved from the database and thus included in the dissemination. There is also a connection into the XML database from the Epic editor, so statistical metadata can also be retrieved via the editor. Statistics Finland chose XML database called eXist to be the database for publications, statistical data and metadata.

25. In consequence, a monthly or quarterly publication written with the Epic publishing editor becomes a document compliant with the CoSSI model and contains all the material of one publication, i.e. text, tables, statistical and document metadata, figures and language versions, in one XML file. These publication originals in XML format are saved in an XML database (eXist), which becomes the publication archive. Published tables and matrices are also saved in XML format into the archive.

26. Before its publishing, a publication in XML format still has to be converted into the format required by the used dissemination channel. For publishing on Statistics Finland’s website, an XML publication is converted fully automatically into HTML and PDF formats. The conversion into HTML format produces a set of HTML pages from one XML publication so that the caption text of the publication forms the start page and the contents listed under it form links to other parts of the publication. The conversion produces sets of HTML

pages in all languages that are present in the XML publication. Besides HTML versions, PDF documents are also produced in each language, and these can be offered to customers as printable versions on the HTML pages. The conversion into PDF format produces one PDF document per each language version in the XML publication.

27. The author of the publication can check the final HTML and PDF versions that will be disseminated prior to their publication. When the publication is ready, the author has at his or her disposal a publishing program for defining when the publication should be released and which files should be published.

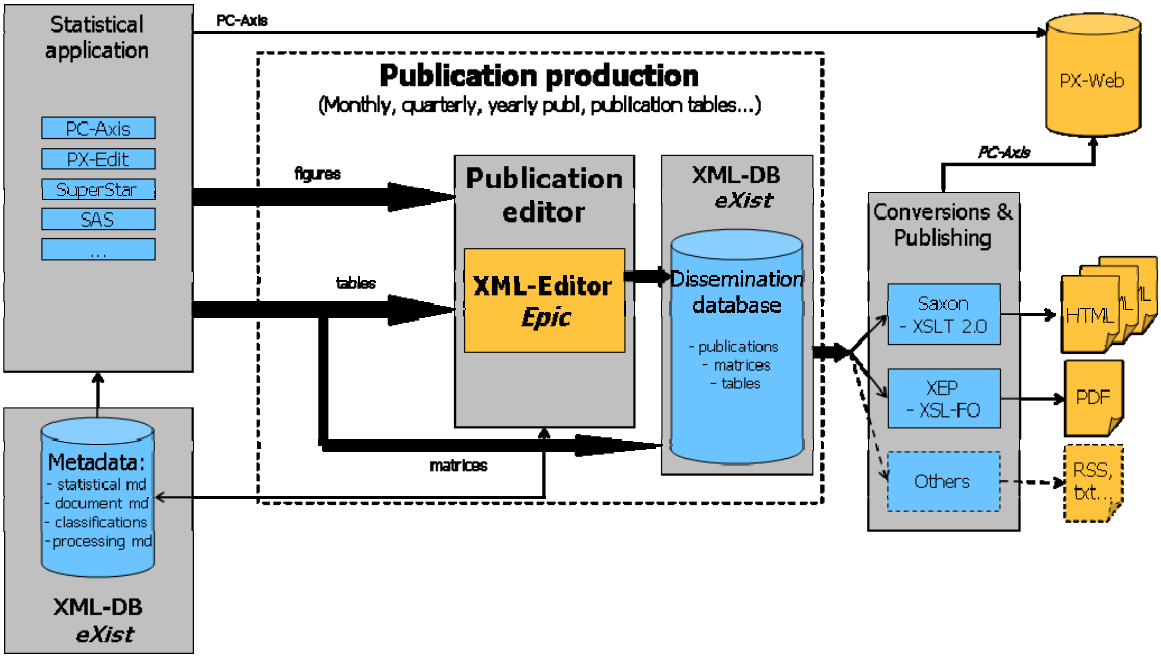

Figure 3. XML based dissemination of statistical data, publications and metadata

IV. EXAMPLES

A. XML database - eXist

28. Open source database called Exist has been chosen as the XML database for statistical publications, statistical data (tables and matrices) and metadata. The database structure has been devided according to the statistics that Statistics Finland produces. Underneath each statistics there are folders for publications, statistical data and metadata of the statistics. In figure 4 there is a view of the database as it appears in the Epic editor database connector interface.

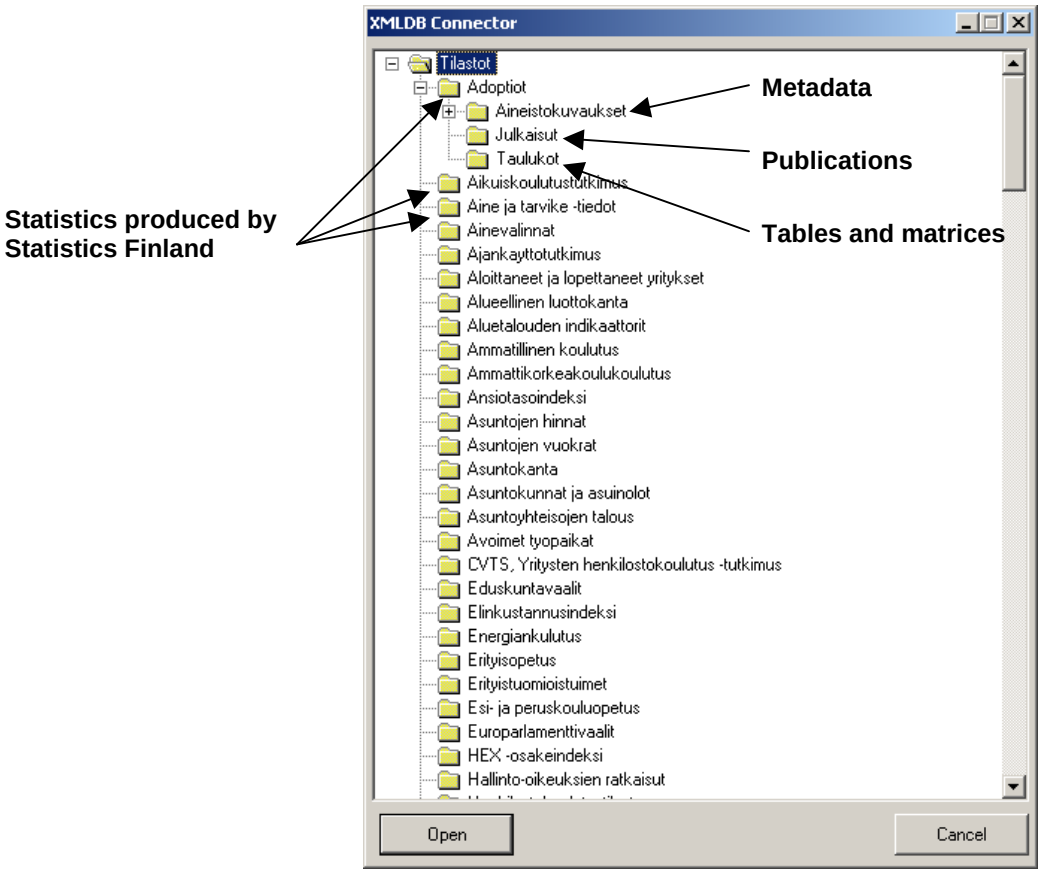

Figure 4. View of the XML database in Epic editor

B. Epic editor

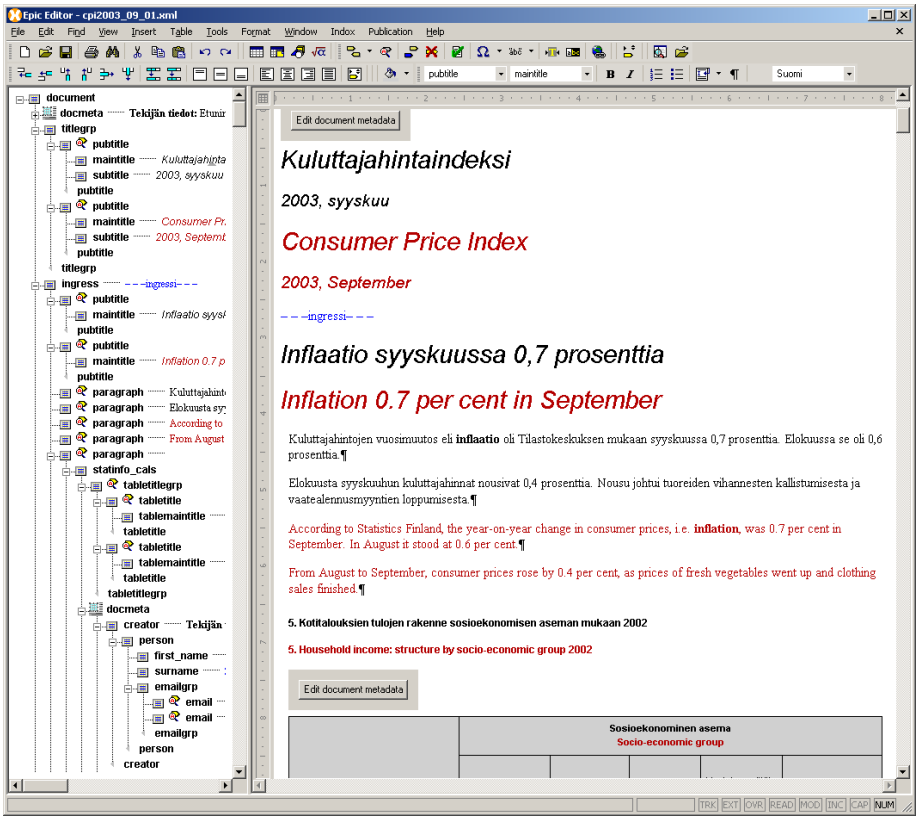

Figure 5. View of a statistical publication in Epic editor

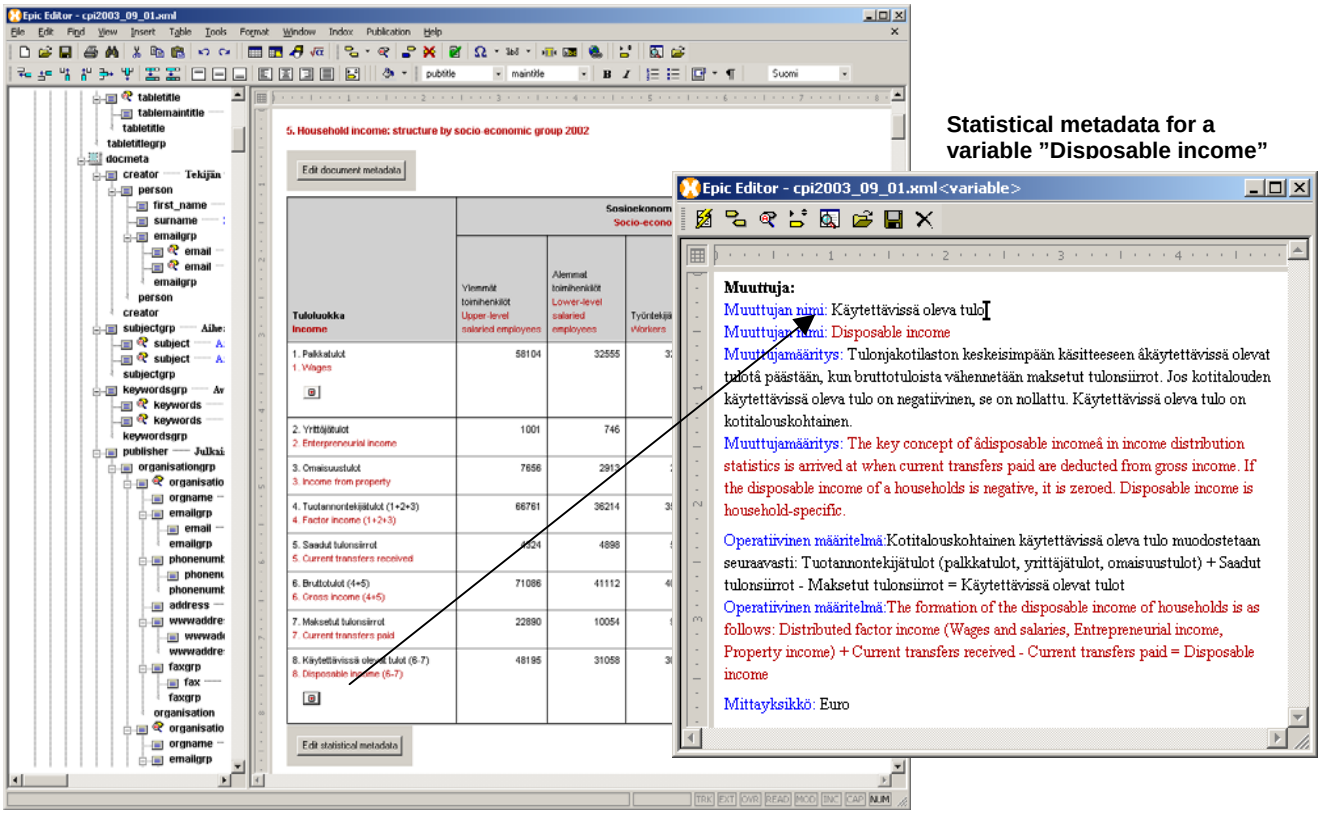

Figure 6. Statistical metadata for a variable in a table

C. HTML output

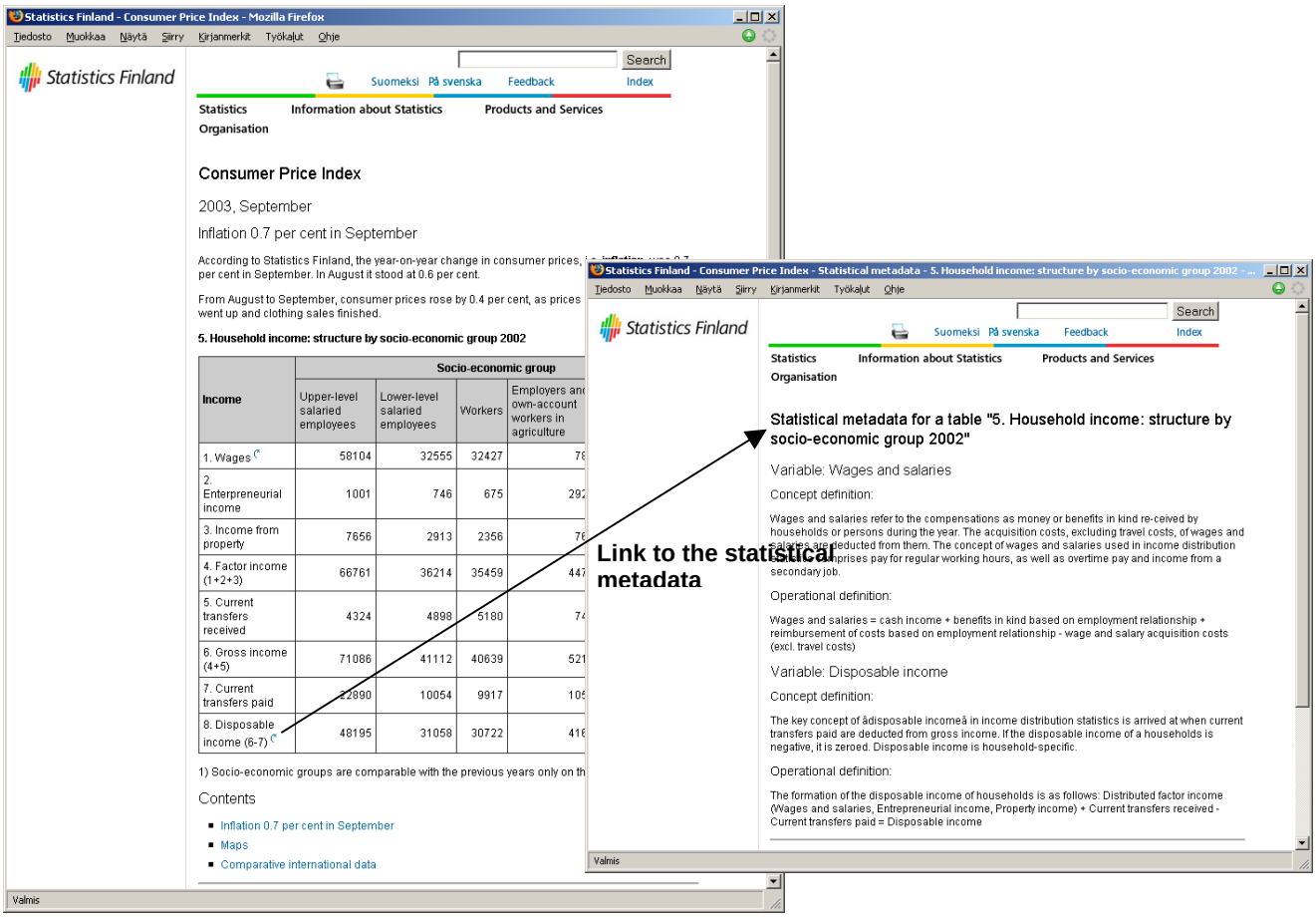

Figure 7. HTML output of a statistical publication with statistical metadata

V. 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. 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., Conceptual Modelling of administrative register information and XML - Taxation metadata as an example. Conference of European Statisticians - Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005

Quality Guidelines for Official Statistics, Statistics Finland, 2002.
Available on the web at: http://www.stat.fi/qualityguidelines