WP.25
ENGLISH ONLY

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

UNECE Work Session on Statistical Dissemination and Communication
(12-14 September 2006, Washington D.C., United States of America)

Topic (iii) How to present metadata

PUBLISHING METADATA WITH DATA - XML BASED DISSEMINATION PROCESS 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.

II. COSSI - COMMON STRUCTURE OF STATISTICAL INFORMATION

3. 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 the situation of the statistical information presented in differing formats.

4. 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

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

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.

5. So the CoSSI model 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, a statistical matrix 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.

III. METADATA PROJECT

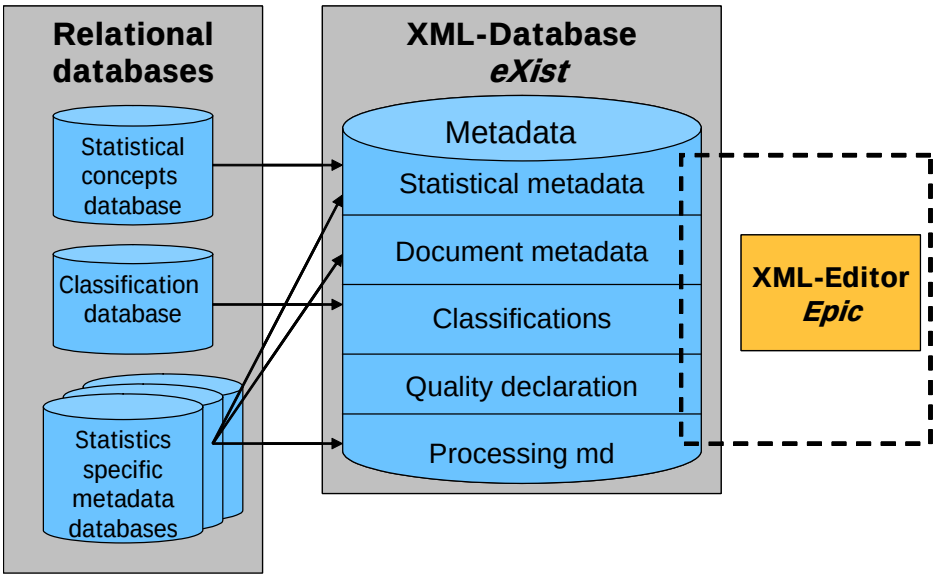
6. Statistics Finland launched a metadata project in spring 2006, with the aims of creating a common display and updating environment for its existing, separate metadata databases. At the moment, metadata are stored at Statistics Finland in the classification database, database of statistical concepts and in various metadata solutions tailored for the specific needs of the agency's statistics departments. These solutions have largely been implemented as relational databases, which are maintained with special software designed for this purpose. The intention is to standardise these separate databases of metadata by creating an XML database that fits the CoSSI model.

7. The task of the project is to ascertain and define how the metadata in the relational databases can automatically be converted to an XML format that fits the CoSSI model and saved in the XML database. A further task of the project is to create a user interface for updating and maintaining the XML format metadata documents in the XML database. The system should also be capable of augmenting fragmented or deficient metadata. The project will also create a comprehensive system of identification codes with which different categories of metadata (statistical, document and process metadata, quality declarations and data descriptions) can be linked to statistical data, tables and publications.

8. eXist database was selected as the XML database. eXist is an open source database that Statistics Finland has tested and piloted in connection with its project relating to the revision of its production of publications. In the aforementioned project the database was principally used for saving tables and publications. This metadata project will extend its use to the metadata side.

9. The planned user interface for the updating of metadata is the Epic editor. By virtue of the project on the revision of publication production, Epic editor will also function as the publishing editor for XML-based publication production and an interface has already been tentatively tailored for it for adding and updating statistical metadata into tables in publications, and for adding metadata to documents. The intention is that the use of the Epic editor will be expanded in the metadata project by improving its user interface relative to metadata and by adding to it new templates that also enable handling of other types of metadata.

10. The intention is to transfer the data into the XML database as XML documents that fit the CoSSI model. Over time, the XML database will, thus, grow into a comprehensive collection of metadata relating to the statistics production of Statistics Finland. In addition, the metadata will then also be in one frame of reference (the CoSSI model) conformant with one model, and in standardised format (XML). As the work progresses, the need to broaden the current metadata definition of the CoSSI model is bound to arise, but any required expansions can be implemented easily and dynamically thanks to the model's properties. Because of the modular structure of the CoSSI model, new components can be added to it and existing definitions in it can be expanded.



11. **Figure 1.** XML database as metadata warehouse

12. An advantage in using the XML technology as the format for metadata, data, tables and publications is that they can all be linked to a common model of information structure - CoSSI. The common model also allows extensive use of standardised tools for data processing. All types of data can be saved in one, common database (eXist XML database), they can be handled and updated using a common interface (Epic editor), and can also be combined and converted using standard XML techniques.

13. The modularity of the CoSSI model also makes it possible to compile a variety of documents for diverse purposes from these data. Tables for publications can be picked out in XML format, statistical metadata describing the data in the tables can be attached to them and the contents of publications and its authors can be described with document metadata. Statistical and document metadata can be attached to a table for publishing it in a database, as well as processing metadata to steer the publishing.

IV. XML-BASED PUBLISHING SYSTEM

14. 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.

15. 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).

16. 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.

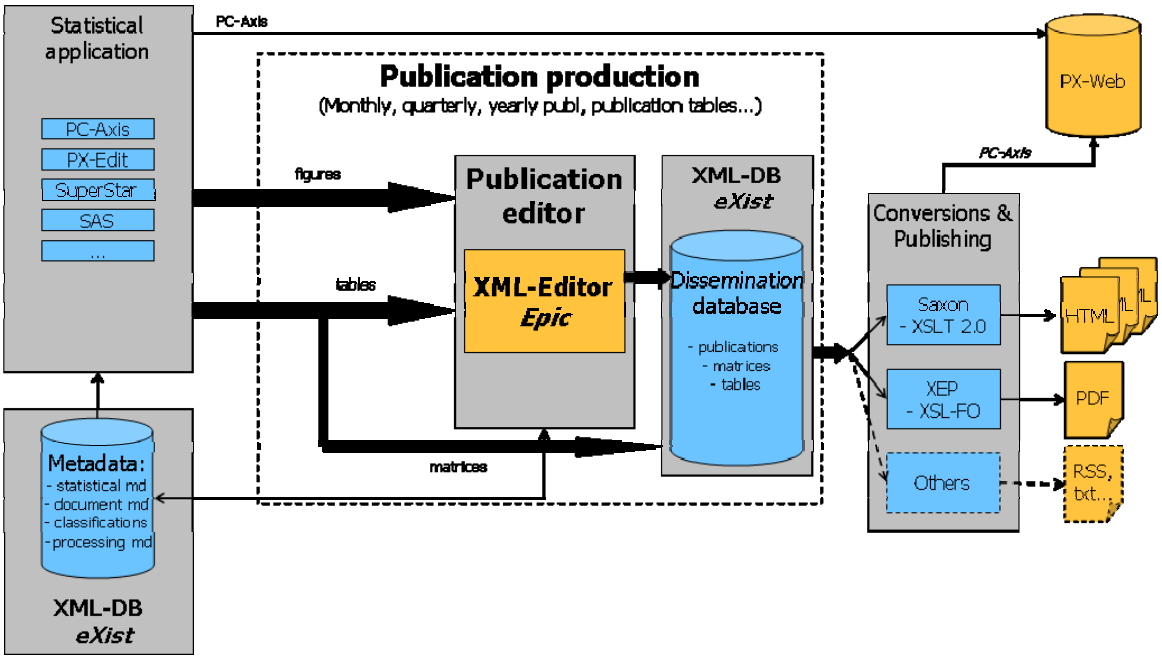
17. An XML database into which statistical metadata compatible with the CoSSI model are saved as XML documents has now been taken into test and piloting use. As tables are made, descriptions of the variables selected for them can be retrieved from the database and thus included in the dissemination. 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.

18. 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.

19. 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.

20. 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.



21.

Figure 2. XML based dissemination of statistical data, publications and metadata

V. EXAMPLES

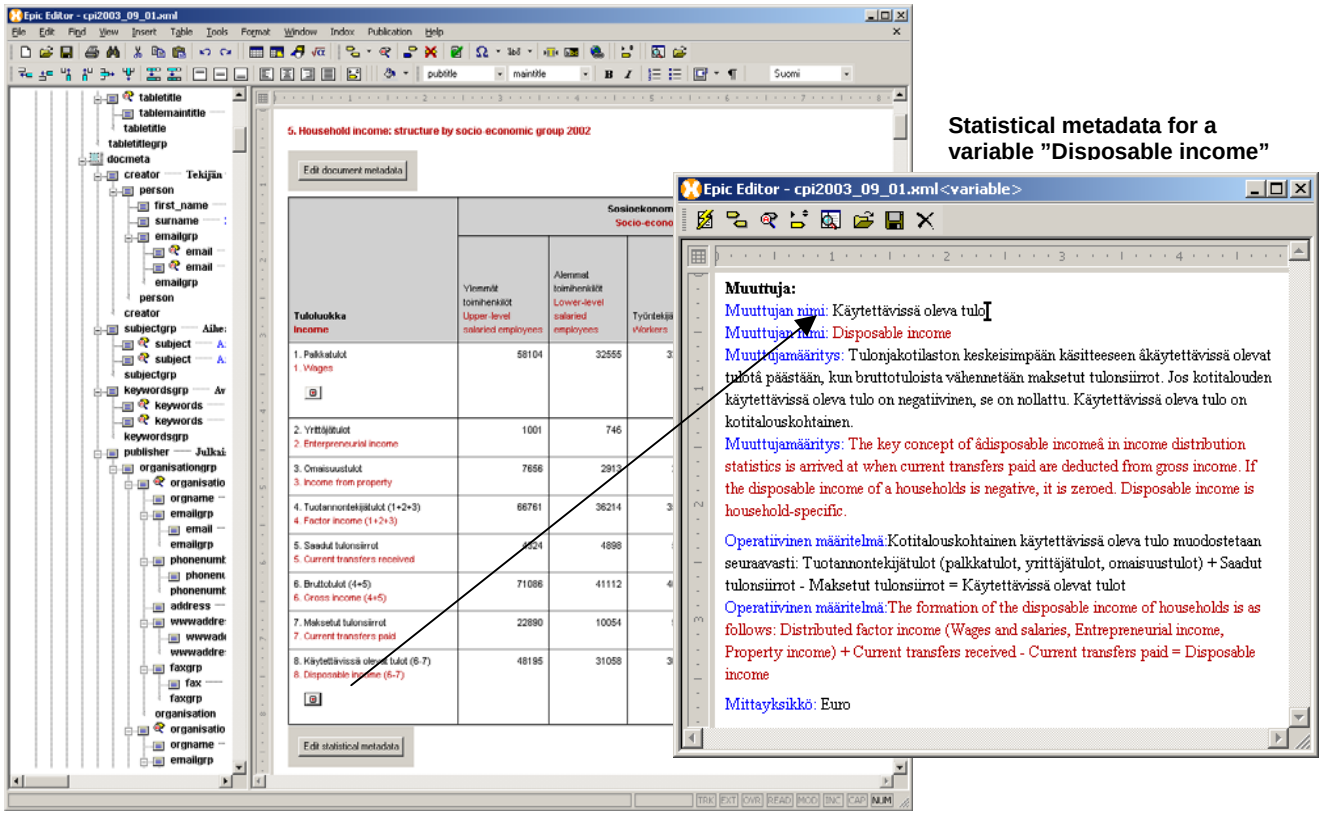


Figure 3. View of a statistical publication and statistical metadata in Epic editor

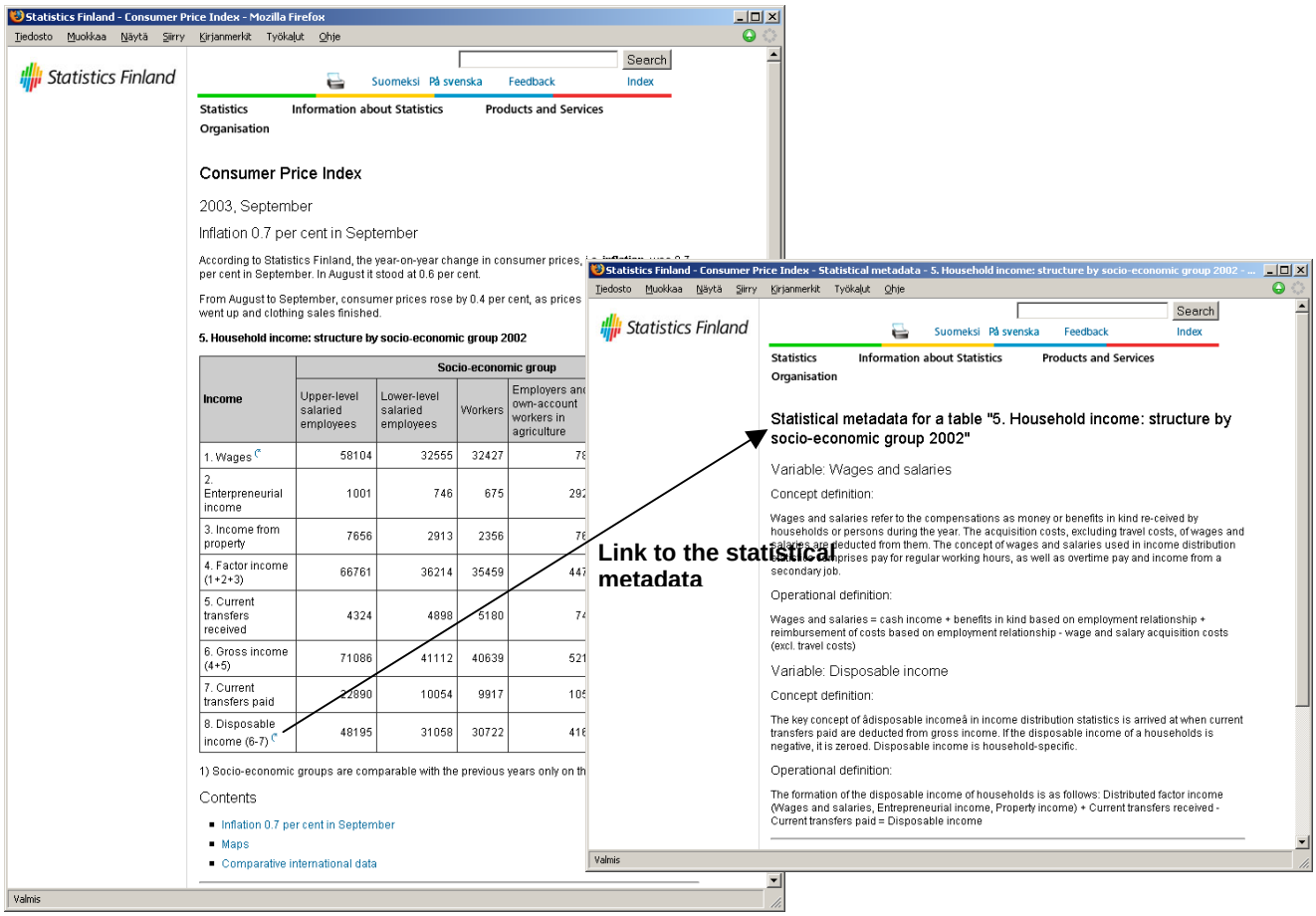


Figure 4. HTML output of a statistical publication with statistical metadata