Distr.
GENERAL

Working Paper No.9
23 April 2007

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**

**Meeting on the Management of Statistical Information Systems (MSIS 2007)**
(Geneva, 8-10 May 2007)

Topic (ii): Statistical information systems architecture

# DEVELOPING STATISTICAL INFORMATION SYSTEMS AND XML INFORMATION TECHNOLOGIES - POSSIBILITIES AND PRACTICABLE SOLUTIONS

**Invited Paper**

Prepared by Heikki Rouhuvirta, Statistics Finland, Finland

## I. INTRODUCTION

1. In the organisation of statistics production the fundamental questions pertaining to information technology have related to – and continue to do so – the question of how statistical data and their contents can be made understood and, on the other hand, to how data warehouses should be organised so that statistical data would be easily available from them in the diverse situations where they are used – whether it be connected with the release of new data or provision of information service from statistics.

2. I will discuss this traditional question-setting from the perspective of one new technology - XML.

## II. SERVICES IN STATISTICS PRODUCTION

### A. Service Oriented Architecture (SOA) in the development of a statistical information system – does it bring fundamentally new perspectives to the examination?

3. In all cases, the exploitation of SOA is based on data – their organisation and management[1]. The solution levels at which data are organised and managed determine largely the ways and possibilities for exploiting SOA. In SOA solutions data are often thought to have been saved in a relational database (RDB), which is assumed to be managed by means of some description or metadata system. It is then presumed that applications, developed for the production of diverse services, whose building is guided by the designing principles of SOA, can be built on this foundation.

[1] See e.g. Lindholm 2004.

4. When we talk about statistical information systems, SOA arouses the following question: what services[2] do we need in the processing of statistical data? How can they, or how should they be defined?

5. The functionality of a typical statistical application can be described in the following way without going into algorithmic details (see Figure 1).

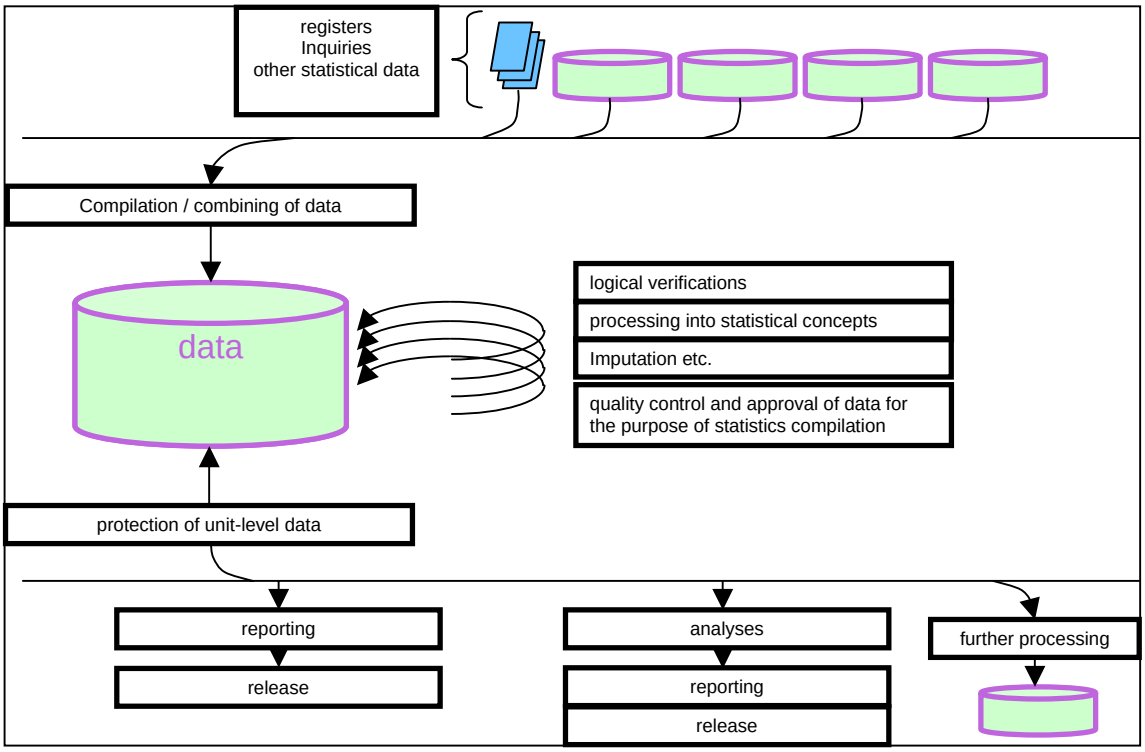


**Figure 1. Typical function specification of a statistical application**

6. In practice, the details of all statistical applications have been built differently and mostly also with deviating application solutions. For example, in 2006 Statistics Finland used over 200 different solutions for individual sets of statistics.

7. The services required in the processing of statistical data can, on the one hand, be specified from the perspective of application development or, on the other hand, by setting out from what is necessary for the processing of statistical data. So that the present status quo of application development would not affect the specifying of the services, it is more meaningful to use the common generic model of statistical data as the setting out point in their specification (see Figure 2).

[2] In this context we exclude administration and projecting from the examination.

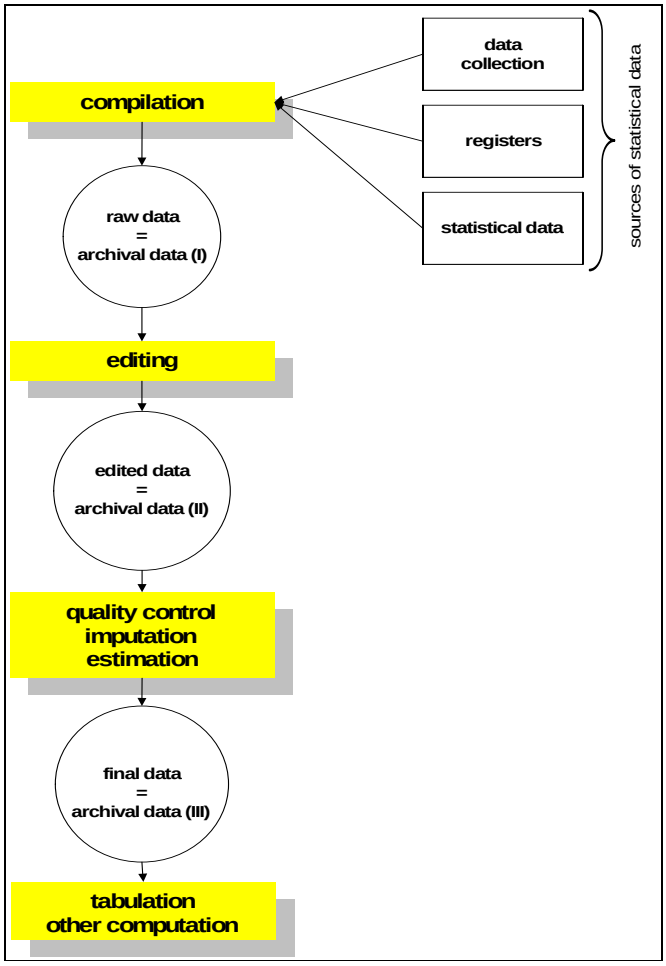


**Figure 2. Methodological processing of statistical data in statistics production**

8. The functions serving the processing of statistical data described in the figure are basic services also in the SOA sense. The special feature of the services is that in one way or another they increase the value added of statistical data. The performed operations – such as transmission or conversion of data – do not in themselves increase the value added of statistical data even though they can be quite laborious as such, requiring intensive input from persons and hardware. However, transmission and conversion may make it easier to introduce the data into use.

9. The basic services improve the quality of statistical data and enhance their content. They can be implemented as applications based on conventional algorithms of statistical mathematics or as copies thereof in a relational database environment. Besides quality control, the core elements in the process are, first and foremost, the understanding of what the data that are being processed mean and what in the manipulative sense can be done to them with statistical methods without affecting their value or meaning.

10. Besides the possibility of editing statistical data, another factor highlighted in the figure (see Figure 2) is that the architecture of statistical systems depends on data warehouses. From the point of the architectures of statistical information systems it is essential that collected and produced statistical data are available for statistical research irrespective of time, place or technical solutions. SOA can then best be understood as an architecture that puts the services produced from a data warehouse on the same starting line. Any statistical calculation can be performed – in other words, service produced – with the data in the warehouse by using the tools best suited for it. With calculation we refer here to data processing from basic description right up to the production of a table, analysis or statistical graphics.

11. In other words, the services are available with which unit-level statistical data can be edited and analysed without compromising the information security requirements on statistical data or having to adapt or reorganise the data in order to be able to perform the specific processing phase concerned.

12. The exploitation of SOA in application development requires a common data model of statistical data. Without it, it will not be possible to utilise all the productivity benefits that come from progress in information technology. What kind of a specification of statistical information would we be looking at then? The basic question in trying to find a solution is how statistical data could be made semantic so that they could be managed by means of a metadata system.

## B. Why XML

13. There are several options that can be used as the setting out point for describing statistical information in information systems. One of these is the ISO 11179 standard[3].

14. ISO 11179 is suitable and is partly also intended for textual describing of data elements. ISO 11179 was born from the need to describe numerical data in relational databases but it cannot create extensive and multi-layered textual descriptions of numerical data. Indeed, the contents of the textual descriptions produced are concise although even quite extensive metadata repository solutions have been implemented based on them[4].

15. Another alternative standard that could be used for descriptions of statistical information is ISO 8879[5].

16. ISO 8879 was born purely for the describing and management of textual data and imposes no restrictions on the extent or the multi-layeredness of the textual data. As a standard ISO 8879 is not coupled with any numerical data or their occurrence in, for example, databases.

17. From the point of descriptions of numerical data the relationship between the ISO 11179 and 8879 standards could be described as follows:

(a) if the value key (what - from where- how: e.g. Value of sales of establishment) is sufficient for describing numerical data, the 11179 is an excellent solution, but
(b) if describing the key requires the forming of a more complex and multi-dimensional domain of concept description, then a mode of description compliant with ISO 8879 is definitely a better solution;
(c) ISO 8879 is a considerably better foundation for the forming of “Electronic Preservation of Data Documentation”.
(d) Correctly defined, ISO 8879 structure contains sufficient information for describing data in accordance with ISO 11179, but not vice versa. There is no other rational standard way of expanding and deepening ISO 11179 descriptions than to transfer the descriptions into the “sphere” of ISO 8879.

18. Today’s realisation of ISO 8879 is XML DTD (Document Type Definition)[6]. XML (eXtensible Markup Language) in itself is a simplified subset of the ISO 8879 standard (SGML)[7]. By contrast, XML

[3] International Standard, ISO/IEC 11179: Information Technology – Specification and standardization of data elements and ISO/IEC 11179-6:1997 Part 6: Registration of data elements; http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=1677.
[4] The MDBD project of Statistics Canada could be mentioned as an example of metadata repository projects (see .e.g. Joannis 2000, 2003)
[5] International Standard, ISO/IEC 8879:1986 Standard Generalised Markup Language (SGML); http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387.

Schema does not in themselves implement the ISO 8879 standard and, in fact, XML is quite often understood as being a programming language or seen as some kind of means of communication. Although XML has been used and applied in program and application development as a specification facilitating easier programming and communication and although these are, indeed, significant features of XML technology, XML is first and foremost something quite other than a programming language or a means of communication[8].

19. The significance of XML is, that it makes possible to describe information in accordance with a standard (ISO 8879) by exploiting the own, natural and logical structures of the information. XML offers a standard method with which diverse structures of information can be flagged and marked, which in turn facilitates easier management of information.

20. It has been regarded as a problem with XML that in marking information structures it defines the syntax but not the semantics, or meanings, for them. This criticism has sometimes been strengthened by demanding that a processing method with which dead structural data (object tree) could be brought to life should be added to the syntax. The structure would become live and significant by attaching a method to it. Of course, attaching scripting to the information structure – in other words equipping the structure with scripts – will enhance the structure with new properties, but is hardly likely to increase the semantic descriptiveness of the structure in the desired manner. Attaching a programmatic description of processing to data does not necessarily improve their semantic understandability. The end result may be a situation that is even more difficult to manage, where the attaching of a programmatic method to the structure joins the program code, data structure and data content even more tightly together into an entity that is even broader but at the same time also more difficult to manage.

21. The basic idea of XML is to create a syntactic specification (marking syntax) with which different structures of information can be distinguished and the contents of the diverse structures can be presented according to a standard without having to produce a separate, own, content-specific and content-dependent specification for each content element. When the structural specification allows simultaneous presentation of varying data contents within just one structure, the structural specification can also be implemented in a semantically sensible manner. Nothing in XML prevents the inclusion of meaning in the structural specification (DTD). In fact, XML specification encourages the introduction of meaning into the structure – especially as the semantics of the structure can be implemented without getting into conflict with a standard (ISO 8879).

22. In respect of statistical information the question concerns the fact that data contents that vary by the object of statistical data (definition of contextual meaning of numerical statistical data, i.e. contextual semantics of statistical data) are presented in the structure of statistical information (semantics of the structure of statistical information, i.e. structural semantics) which for its part reveals what matter or description related to statistical data is concerned. Thus, from the perspective of statistical information the structural specification of statistical data is semantically significant.

23. The self-evident and semantically ready-defined elements of a semantically significant structural specification of statistical data include such general concepts of statistical data as table, table heading, data, observation or statistical unit, variable, category value, etc. In statistical information they are clear and easily interpretable elements possessing a semantic meaning. By using them, semantic relevance can be given to the structural specification of statistical data. A semantically relevant structure helps the statistician and the user of statistics to control the correctness of contents.

24. The end product can be presented as a (three-tiered) model of statistical information in which

---

[6] For basic definitions, see W3C at http://www.w3.org/XML/.
[7] For brief introduction to XML, see e.g. http://en.Wikipewdia.org/wiki/xml.
[8] What is usually meant by this is that because messages in XML format are not dependent on tools or architecture they offer the possibility of integration between entirely different types of applications.

*data = value + structure (implemented as semantically meaningful) + semantic description.*

25. Statistical information (data) without a semantically meaningful structure is ineffective in its realisation. On the other hand, exploitation of all levels of the presented data model makes the solution more informative and easier to control from the point of statistics production[9].

26. Thus, it is no so much a question of deficiencies or problems in the marking standard than the way the standard is implemented. An example of a semantically meaningful structural defining of statistical data is CoSSI[10].

## III. USE OF XML IN STATISTICS PRODUCTION[11]

### A. Specification of a statistical data

27. To standardise the use of XML in the production of statistics we have produced an expansive structural specification of statistical information within which the information content of statistical data are described in a standard manner. The aim has been to produce a mode of description that people can understand and from which descriptive information can be exploited in applications as such. It is also the aim that the contents of the descriptions of statistical data would be suited for a variety of purposes of use, and be adequate and understandable to the users of statistics themselves.

28. The structural specification of statistical information covers both diverse data contents of the information and various ways of organising statistical information (see Figure 3).

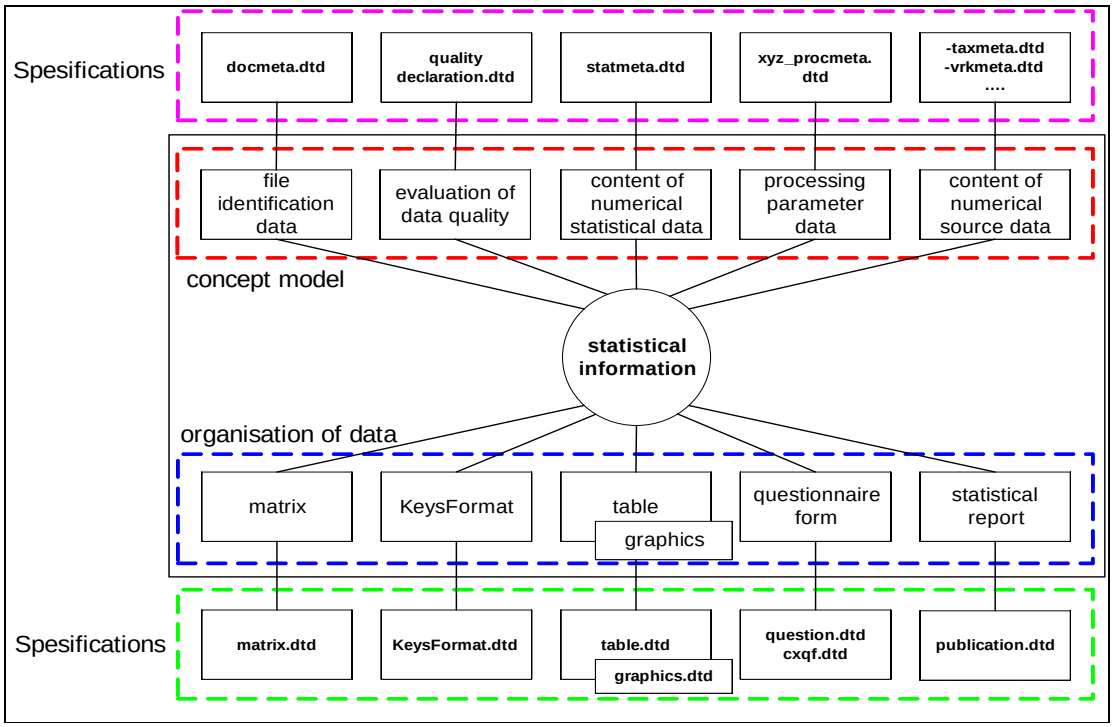


**Figure 3. Structural specification of statistical information**

[9] Lack of the second level (semantic structural specification) makes the solution an ineffective description platform. A good example is SDXM in which the ineffectiveness factor that is crucial from the point of presentation of statistical data is connected precisely with the fact that the model lacks a semantically meaningful structure.

[10] CoSSI, see Rouhuvirta – Lehtinen 2003; http://www.stat.fi/cossi

[11] This examination focuses on statistical processing of data only excluding administrative services and project management.

29. Our intention at Statistics Finland is to apply this specification at different stages of statistics production so that statistical data themselves do not need to be respecified or redescribed as they move from one production stage to another, but as the content of the data remain unchanged they will remain automatically available in all the production stages the data are being processed in one way or another (see Figure 4).

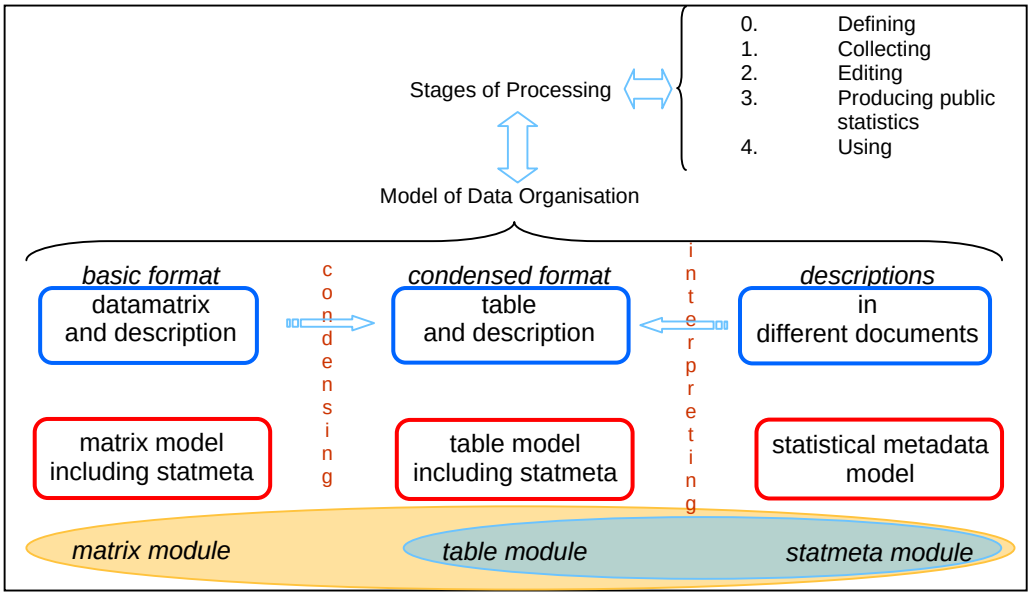


**Figure 4. Statistics production and ways of organising and content descriptions of statistical data**

30. Next, I will illustrate how XML solutions are exploited in different stages of production.

## B. Data description and its use (data descriptions and data)

31. A data description contains descriptive information on the content and methodology relating to all variables in one set of statistics. In addition to this, a quality description can also be added to the data.

32. The data descriptions are saved into an XML database (eXist). The structure of the database available to the statistics departments is presented in the adjacent figure.

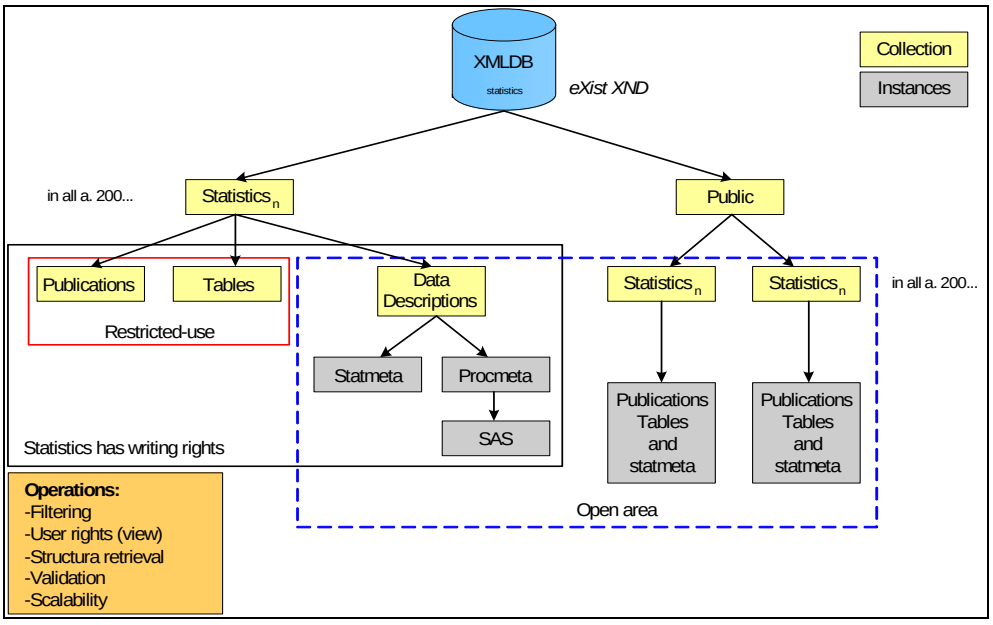

**Figure 5. Data descriptions of statistics and statistical information in the XML database environment**[12]

33. Use of the data descriptions in application-level solutions (SAS and tabulation) is presented in Figure 6.

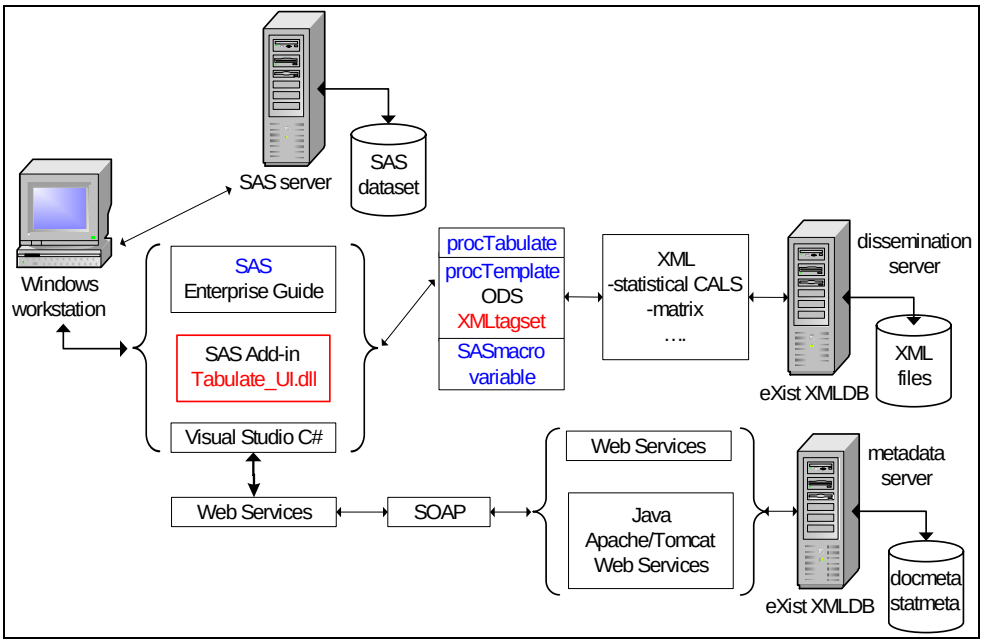


**Figure 6. Application architecture of tabulation environment in SAS**

## C. Data editing (data and data description and management of new data)

34. Descriptions of derived variables formed for statistical data and the procedure for handling them in SAS environment are presented in the adjacent figure (see Figure 7).

[12] See actual list of statistics in Statstics Finland at http://www.stat.fi/til/abcd_en.html.

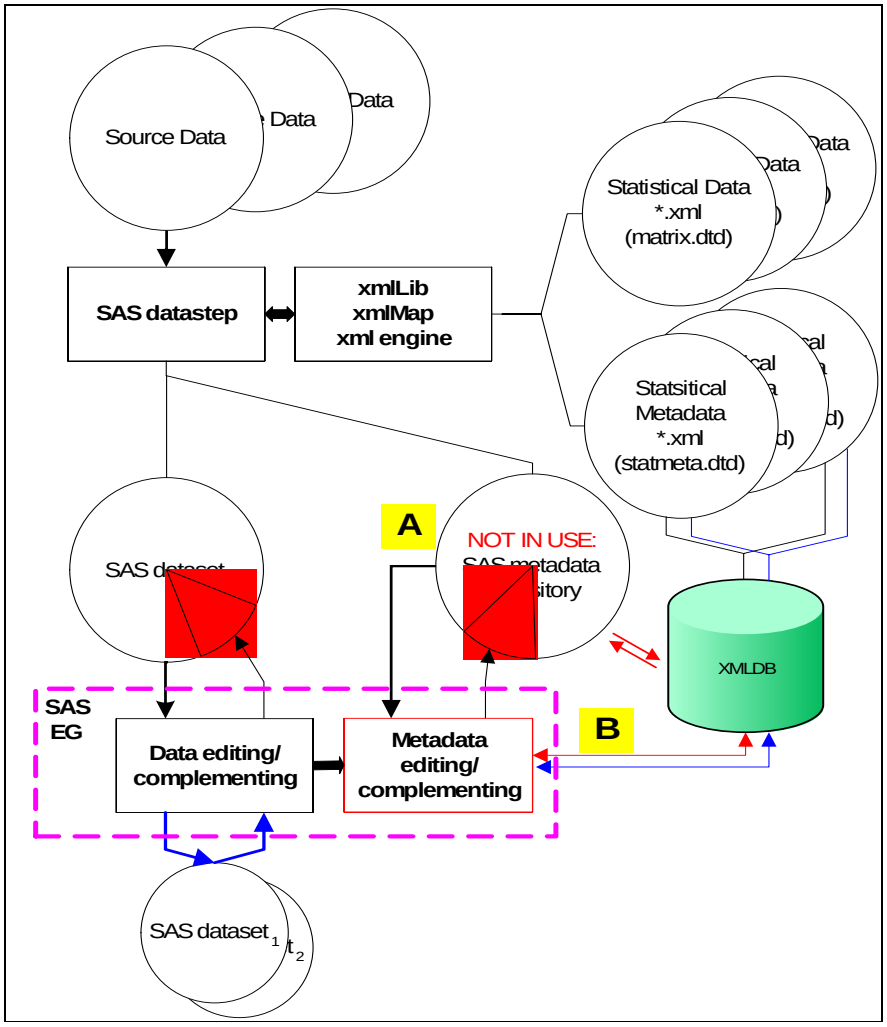


**Figure 7. Data and statistical metadata managing in SAS data editing procedure//process**

35. When new, derived variables are added to data, their content and the methodology by which they are formed must be described. The description is produced in the editing application environment in an application dialog tailored for this purpose and with SAS EG Client or the XML editor used at Statistics Finland, and saved as part of the data description in the XML database (see choice B in Figure 7.). The XML database also allows the use of browser interfaces.

## D. Backup, archiving and input of relational databases into mathematical statistical applications

36. An XML specification of statistical data in matrix format (see Figure 8), which contains the data, file description and statistical metadata, functions as a unifying file format that can be used:

(a) as a backup file of relational databases;
(b) as a transfer file from relational databases to mathematical statistical applications (editing, analysing and tabulation applications);
(c) As an archiving format.

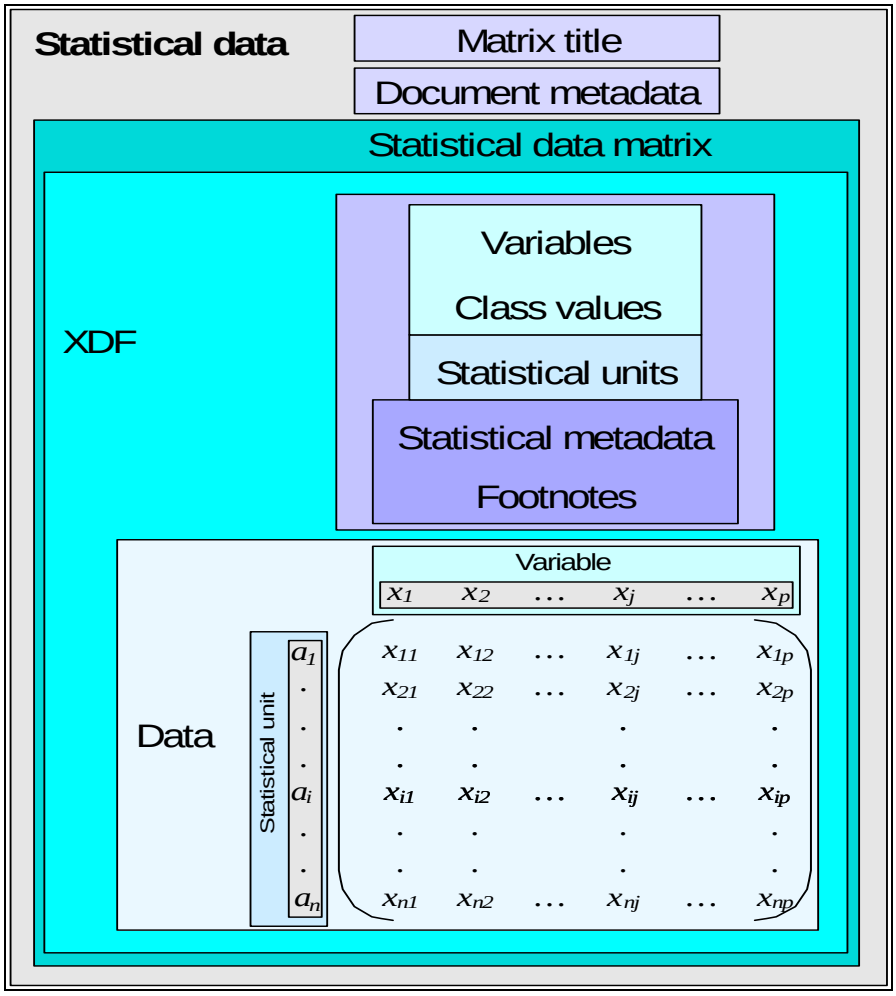


**Figure 8. Logical schema of an XML file compliant with matrix specification**

37. At the application level the SQL/XML access creates good preconditions for the automation of the backup functions of relational databases. The matrix specification of statistical information, in turn, makes the backup files independent of the version of the database application and at the same time suitable for archiving. In the matrix format the file also includes the metadata retrieved from the XML database. In addition, the same file format can also be used for input into e.g. SAS.

38. Application-specific data are always application-specific and their maintenance, such as formal control of numerical formats, is always included as a part of output or input application.

### D. Data warehouses, release and dissemination

39. Tables are produced[13] as electronic XML originals compliant with tabulation specifications. The originals can be attached to a publication in the XML editor or published independently. The dissemination application takes care of the publishing and printing of tables in the desired format, such as html or pdf. The process is described in Figure 9.

[13] The basic tools for the production of tables are SAS and SuperStar. We can produce XML tables compliant with the Statistical CALS specification (see Rouhuvirta – Lehtinen 2003) with both.

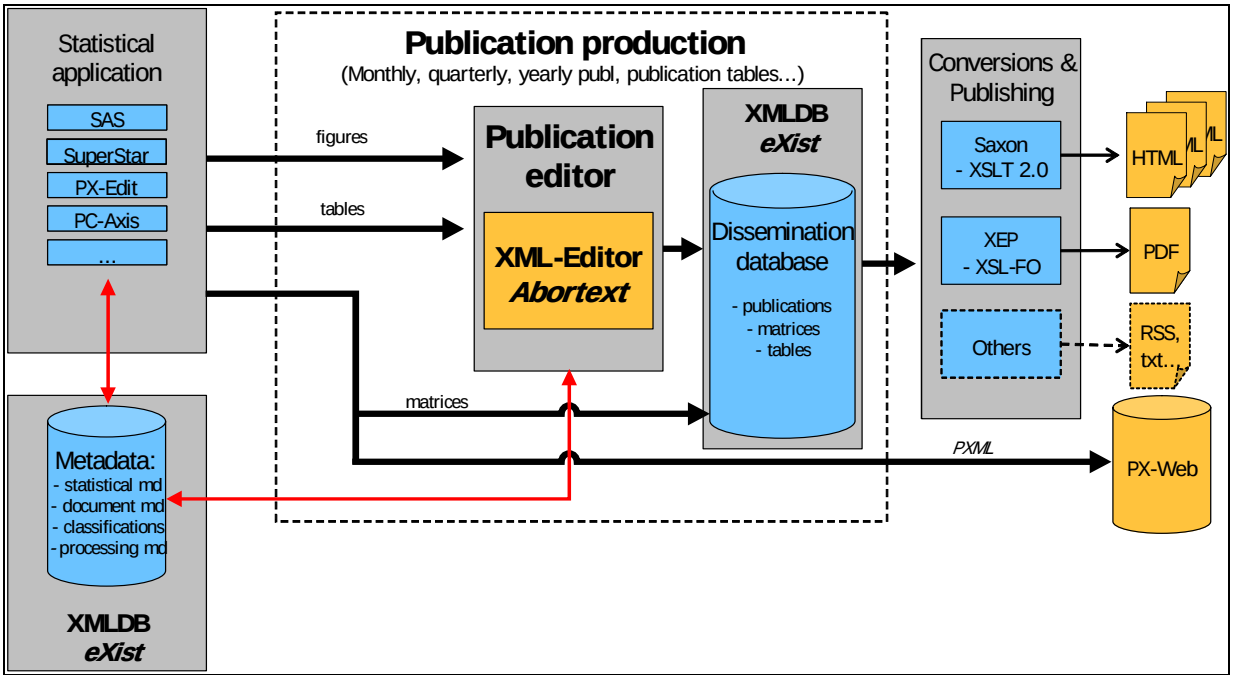


**Figure 9. Production and dissemination of tables as part of the publishing process**

40. The table originals are saved into the XML database.

## E. Retrieval of statistical data and information services

41. Structured full text search (structured free text search) is used in the retrieval of statistical data both during statistics production and in the provision of information service subsequent to dissemination. The search utilises all the textual information of statistical data that has been saved during their production process into the metadata descriptions contained in the CoSSI model. A structured search means that the structural elements of the structural specification of statistical information, and their semantic meanings, can be used in the generation of search terms.

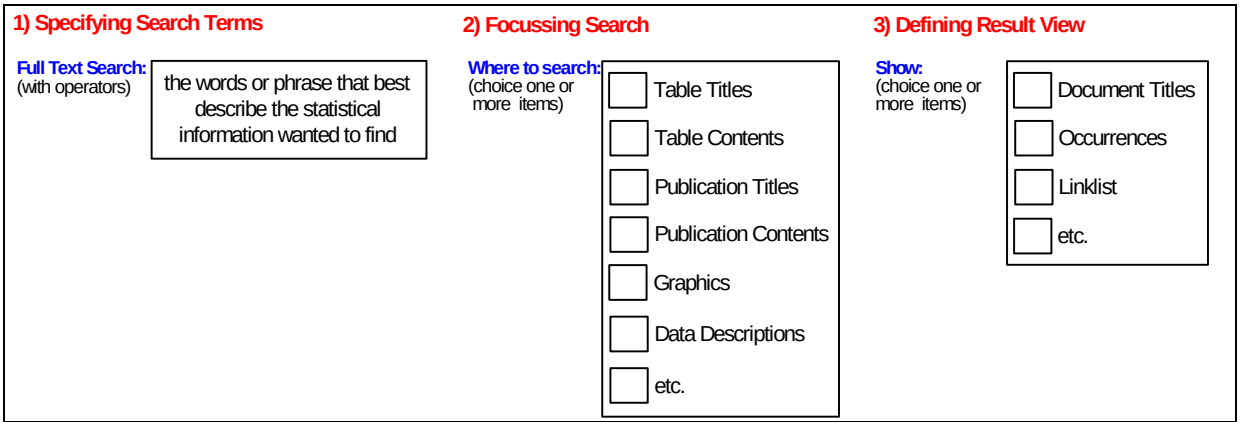


**Figure 10. Basic elements of user interface for retrieval of statistical information including statistical data**

42. The background applications for the search function are the XML database and its indexing application. The same indexing solution will in future also be used in the search service functions on the Internet pages.

43. The data (information) that can be searched will allow the user to search both by the content (variables, classifications, tables, etc.) and by the descriptions of the content and quality of found numerical data.

## IV. CONCLUSIONS XML BASED SERVICE ENVIRONMENT IN STATISTICS PRODUCTION

44. The statistics production solution briefly described above gives indications of the kinds of services that could be produced from a statistical information system in future, both for statisticians and the users of statistical data. The foundation (for statistics production) is an XML-based information architecture and standard applications exploiting it.

45. Basing the implementation of the information architecture on XML allows utilisation of standard and standard-like specifications, but the special characteristics of statistical information should be taken into consideration in their application and implementation. If, for instance, the possibilities of a semantic structural specification are not exploited in the structural analysis and the final structure of statistical data, from the point of information management the solutions become complicated, on the one hand, and ineffective in practice, on the other. From the perspective of application development, it seems especially important that the information architecture itself does not contain application-specific data specifications, because we are unlikely to see a situation where we would have just one monolithic application for both statistics production and information service provision.

46. A semantically relevant structure helps the statistician and the user of statistics to control the correctness of contents.

### *References*

International Standard, ISO/IEC 11179: Information Technology – Specification and standardization of data elements and ISO/IEC 11179-6:1997 Part 6: Registration of data elements; http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=1677

International Standard, ISO/IEC 8879:1986 Standard Generalised Markup Language (SGML); http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387

Johanis, P., Metadata at Statistics Canada, Canadian Metadata Forum 2003, September 19-20, 2003; http://www.collectionscanada.ca/metaforum/014005-03219-e.html

Johanis, P., Statistics Canada's Integrated Metadabase – Current Status and Future Plans, UN/ECE Work Session on Statistical Metadata, 28-30 Nov 2000; http://www.unece.org/stats/documents/2000/11/metis/3.e.pdf

Lindholm, M., SOA - Mistä Kysymys? - Nykytila, mahdollisuuksia ja kokemuksia Capgemini Finland 2004; http://www.jhs-suositukset.fi/intermin/hankkeet/jhs/home.nsf/files/jhs-seminaari2006-04-26lindholm/$file/jhs-seminaari2006-04-26lindholm.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

The World Wide Web Consortium, W3C Recommendation: Extensible Markup Language (XML) 1.0 (Fourth Edition); http://www.w3.org/TR/2006/REC-xml-20060816/

---