STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 19
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Electronic Data Reporting
(Geneva, Switzerland, 13-15 February 2002)

Topic (iii): Metadata, conceptual models and standards

METADATA-BASED SYSTEMS AND XML-BASED DATA FORMATS IN THE PRODUCTION OF STATISTICS IN GERMANY

Submitted by the Federal Statistical Office (DESTATIS), Germany[1]

Contributed paper

ABSTRACT

The Federal Statistical Office is currently developing and implementing XML-based document types for statistical tables and data that will make it possible to handle and relate data and associated metadata in a consistent manner throughout the complete process of statistical production. Major long-term objectives include application and data integration, quality assurance and improvement, and archiving data in an application-neutral format.

In Germany, public surveys are conducted by the Federal Statistical Office in cooperation with the 16 Statistical Offices of the German states ("*Länder*"). The German Statistical Offices have a long tradition of developing common software products and establishing common standards. They have decided to use XML-based document types for the whole of the statistical production from data reporting to publication. Initial steps include the development of a document type for tables (*TabML®*, *Table Markup Language*) and of another one for statistical data (*DatML®, Data Markup Language*).

The most important tools in use at the German Statistical Offices are SPLV®, a 4GL programming language for statistical purposes, STATSPEZ®, a metadata-based tool enabling non-programmers to specify and produce tables with automatically generated SPLV programs, and GENESIS®, a statistical information system. In addition, off-the-shelf products like SAS are used. All the metadata of an SPLV program can be extracted automatically. STATSPEZ stores metadata in the form of reusable objects. Shared STATSPEZ metadata data bases are being set up by the German Statistical Offices to form a common metadata repository. SPLV and STATSPEZ alone provide for about 90% of the regular tabulation throughout all Statistical Offices. As a consequence, ongoing and upcoming efforts to introduce XML document types are centred upon them.

TabML defines the Layout Format for the visualization and presentation of tables and the Matrix Format with a complete set of metadata for in-memory tables ("matrices"). DatML is designed for describing statistical data and for supporting the transmission of data from questionnaires, data editing and statistical non-disclosure. Both document types are implemented in the form of modular, fully parameterized XML DTDs. They share a common set of element types for document-related – mainly statistical – metadata defined in another set of DTD modules, the *Core Tag Set* (CTS). TabML, DatML and the Core Tag Set are free for anyone who wishes to use them.

[1] Prepared by Michael Shäfer.

The STATSPEZ functionalities will be extended in order to turn it into a metadata administration tool for all stages of the statistical production. It will support the processing of paper questionnaires and be used to generate electronic questionnaires. For data editing, STATSPEZ will allow the generation of program code in various languages (SPLV, Java, Blaise). STATSPEZ will provide means of administering and processing data and metadata consistently from data reporting to tabulation.

The XML-based document types TabML and DatML - and probably others - are developed for data and application integration, and provide a set of document types that allow to store metadata consistently. They will be implemented into STATSPEZ and SPLV for the purposes of data editing and tabulation and into the GENESIS information system and third-party tools to make use of the metadata in the areas of information retrieval, statistical analysis and publication.

I. INTRODUCTION

I.1 German Statistical Offices

1. In Germany, public surveys are conducted by the Federal Statistical Office in co-operation with the 16 Statistical Offices of the German states (“Länder”). Although being legally and organisationally independent, these institutions have a long tradition of co-operation in various fields. In the field of data processing, they share their programming work load, establish common hardware and software standards, and develop and implement software and procedures based on these standards, which, in turn, enable further co-operation in spite of the growing number of hardware platforms, operating systems and tools in use. Common metadata-based systems - mainly for data evaluation - have been developed and are productive.

2. The majority of the common standards apply to software development and third-party software tools. They cover more or less isolated branches or steps of data processing and the statistical production, but hardly deal with the issues of metadata and of data and application integration. In recent times, deficiencies have come up in these areas and now prove major obstacles on the way to higher quality and efficiency of the statistical production.

3. National Statistical Offices of other countries and other national and international institutions have made considerable progress in standards for data and metadata exchange, leading to results like GESMES-cb, RDRMES, eQuest and IQML. However, even without standardization on a larger scale, the use of XML document types is a step in the right direction and goes a good part of the way to future standards. Therefore, the Statistical Offices have decided to use XML-based document types for the whole of the statistical production from data reporting to publication. Initial steps include the development of a document type for tables (TabML® – Table Markup Language) and of another one for statistical data (DatML®, Data Markup Language), plus adequate Java-based tools. This decision is seen as a necessary and consistent measure to pave the way for up-to-date tools and methods and provide a long-term prospect for the quality, efficiency and reliability of the statistical production.

II. CURRENT STATE OF THE STATISTICAL PRODUCTION PROCESS

II.1 Standardized Tools

4. The common sofware standards of the German Statistical Offices include a variety of tools, mainly for the tabulation and evaluation of statistical data. In addition, non-standard tools are in use at the Federal Statistical Office (and others) to cover other areas such as data reporting (for example, W3STAT, an application that allows data reporting for foreign trade statistics via the Internet). However, many of the non-standard tools are designed for specific statistics, whereas most standard tools have in common that they can be used generally and have been designed and developed under direction of and with contributions from one or more of the Statistical Offices. Among the last-mentioned, the most important – measured by their coverage of the statistical production – are SPLV®, a 4GL programming language

for statistical purposes, STATSPEZ®, a metadata-based tool enabling non-programmers to specify and generate tables, and GENESIS®, a statistical information system. The following sub-sections provide information about these tools in detail. In addition, off-the-shelf products like SAS are used.

5. Altogether, a large number of tools are in use, but the core of the statistical production - data evaluation and tabulation - almost completely relies on SPLV and STATSPEZ (which is based on SPLV). They provide for about 90% of the regular tabulation throughout all Statistical Offices, and - quite naturally - ongoing and upcoming efforts to introduce XML document types are centred upon them.

II.2 SPLV – a 4GL Programming Language

6. SPLV is a strongly structured, self-documenting and easily maintainable 4GL programming language for statistical purposes. An SPLV program consists of a number of so-called building blocks performing specific functions on the data within their scope. Typical functions are read, write, sort and synchronize data streams, build tables, output tables in a printable format and process hierarchically organized data. Other building blocks provide access to in-memory or external reference data or allow the specification of useful language elements that contain metadata, such as a record structure or a statistical or hierarchical attribute. Data streams connect the building blocks, and the logic controlling their flow is generated by the SPLV compiler. SPLV is productive since the end of the 1980s. Regular annual updates are made to extend its functionality and adapt the language to user requirements. Not being a third-party tool, SPLV can be extended as the need develops. It is fully integrated into STATSPEZ (see below), and compilers and run-time environments are available for IBM OS390, Fuijitsu-Siemens BS2000, Sun Solaris and Microsoft Windows NT.

7. An SPLV program contains and produces various types of metadata that can be useful in statistical (i.e. properties or attributes used in constructing the table matrix), organisational or technical contexts. Most of the metadata of an SPLV program already exists at compile time within the program code, either explicitly - serving to facilitate programming and constructing the program - or in the form of "hidden" metadata that can be derived from the context in which a language construct appears, etc. This kind of metadata is static, as opposed to the variable metadata created at run time, such as the program execution time and environment, and the names of data sets. Run-time metadata may also be of statistical value, i.e. the number of records (representing statistical units) processed from a data set.

8. SPLV programs use a specific XML document type called SPLVML to output both metadata and data. Documents of that type can then be converted into documents of other, arbitrary types like TabML and DatML. Although necessitating an extra conversion step, this is a useful technique because it detaches compiler from document type. SPLVML is planned to support three kinds of data streams: printable tables (to be converted into the TabML Layout Format), in-memory tables (to be converted into the TabML Matrix Format) and "ordinary" sequential or hierarchical record sets (to be converted into DatML). Currently, printable tables are supported.

9. All the metadata of an SPLV program can be extracted automatically. Existing programs require no or very few changes to the program code, and a simple recompilation will do in most cases.

II.3 STATSPEZ – Specifying and Generating Tables - and More

10. STATSPEZ is a metadata-based tool for the comfortable, GUI-supported specification of statistical tables, allowing ad-hoc aggregation and presentation of statistical data. It is a recent development that went productive in 1998 and has originally been designed for the use in statistical departments as an easy-to-use tool for the generation of SPLV programs. Further STATSPEZ features include an FTP-based file transfer method, a mechanism for defining and controlling automated production steps, support of ASCII and EBCDIC data types, automatic conversion of file and data format and encoding, and the possibility of cross-platform development. STATSPEZ is only available for Windows NT 4.0.

11. All metadata is stored in a data base in the form of reusable objects, holding the information needed to automatically generate and run SPLV programs, so programming skills are not required. Example objects are data set descriptions, i.e. records and fields with their data types, statistical attributes and properties, table structure and table layout. STATSPEZ offers sophisticated methods to arrange these objects hierarchically to meet organisational requirements and restrict both visibility of and access to them. The metadata data base may reside on the local Windows NT system or on a server.

12. Shared metadata data bases are being set up by the German Statistical Offices to form a common metadata repository. All hierarchical levels can be declared private or public, avoiding setting up separate data bases for private and shared objects. An administration server takes care of changes to the metadata data bases and keeps them in a consistent state.

13. At the time being, STATSPEZ is mainly a tabulation tool, but steps are being taken to extend its functionality onto the areas of data reporting and data editing, turning it into a powerful tool for metadata administration at all stages of the statistical production.

II.4 GENESIS – an Information System

14. GENESIS is an information system designed for maintaining and evaluating a statistical data and metadata data base. Its core functions are data and metadata administration, import and export of data and metadata, conducting searches in the metadata data base, dynamic definition and compilation of tables, and display or print the respective results. GENESIS is available as an integrated system on IBM OS390, Fuijitsu-Siemens BS2000 and Sun Solaris platforms and as a client/server system with one of the before-mentioned systems acting as a server for MS-DOS or Windows-based client systems.

15. The GENESIS data base not only holds statistical values and the associated metadata, but comes with a sophisticated metadata retrieval system. The metadata consists of information about surveys, statistical attributes and properties with their values plus rules for constructing derived attributes and properties. A thesaurus facilitates searching the metadata data base.

16. GENESIS stores arbitrarily ordered statistical data and allows its (re-)combination. It does not require data to be ordered by a specific primary key. For example, data can be deeply structured and ordered regionally, it can be a time series or express correlations as in migration statistics.

17. The logical structure GENESIS uses to store its data can best be thought of as a cube with n dimensions or axes. Accessing data in such a cube requires specifying a survey, specifying the dimensions or axes by giving a statistical context , i.e. a classification, a regional and a time context, and selecting the statistical values.

III. TABML AND DATML

III.1 Advantages expected of XML-based Document Types

18. Most of the advantages expected of XML-based, self-describing documents are due to the fact that such documents are free from proprietary data formats and not only allow the exchange of data but of metadata as well. Traditionally, the metadata necessary to process a file resides in the application and is therefore inadequate to deal with arbitrary file structures, not even with those of the same type. Making metadata part of the document kind of disconnects applications from data. This leads to generic processing where an application is designed for processing a specific document type and can deal with all instead of one or a few of document instances, provided they conform to the document type. As a consequence, this reduces the number of applications, especially for post-processing steps, and lessens the strain on application development, saving efforts and costs. It also makes application integration easier and more worthwile, because implementation efforts concentrate on a relatively small number of

applications, with each having a potentially very much higher number of documents and amount of data eligible for processing.

19. Data integration is another hoped-for advantage from a statistical and organisational point of view. While application integration connects applications using the same document type, data integration brings together documents of different types and, for example, allows to associate a table with the input data used to produce it, or to retrieve all documents associated with a specific survey and a reference period.

20. Quality assurance and improvement can be reached in two ways: enhancing the accurary of data processing with metadata, and generic processing. Lessening the number of applications through generic processing comes along with fewer sources of possible errors and long-lasting, reliable software. On the other hand, errors in generic processing potentially affect much more data.

21. Non-proprietary, readable data formats and self-identifying documents help improve archiving, because the risk of not being able to retrieve, read and understand them is no longer an issue of the data itself, but reduced to the problem of keeping it on an accessible storage medium. Additionally, it facilitates archiving self-describing data from aspects of its contents or in context with other data.

22. As for statistical tables, most of the current production ends up in file formats that are good enough for print and display, and very often unfitted for generic processing, or in other words, end-products. TabML shall eliminate these restrictions and help transform tables into an easily processible data source.

23. DatML shall describe sequential and hierarchical record structures and support data editing and data integration with TabML and other document types.

III.2 Current Support of TabML and DatML

24. The current release of the XML vocabulary TabML defines the so-called Layout Format (see below). It is supported by SPLV and STATSPEZ, meaning that those tools can output tables in that format and post-process them using built-in functions. In addition, first steps have been taken to use TabML as an output format for GENESIS and as both an input and output format for SAS applications. The first release of DatML is scheduled for mid 2002.

25. TabML 1.0 is implemented in the form of an XML DTD. DatML will be implemented in the same way. Future versions of these document types may be defined using a different schema language. To facilitate maintenance and enable adaptability to user requirements, the TabML DTD is modular and fully parameterized. This makes it possible for users to easily create a new DTD that better serves their needs by building a subset of or extending the original one.

26. Since both TabML and DatML are designed for the area of statistics, it is natural that they can contain the same kind of metadata, i.e. information about a survey. To ensure that such metadata is stored in documents of both (and possibly future) types in a consistent manner, another set of DTD modules, the Core Tag Set (CTS), has been created that defines element types and attributes for statistical and other, more general purposes as well. The TabML and the future DatML DTD reuse identical element type and attribute declarations from the Core Tag Set (CTS) and thus store the respective metadata consistently.

27. TabML and DatML and the Core Tag Set are free for anyone who wishes to use them. Some general-use software, especially converters, is offered for free as well. All element and attribute names and terminal names appearing as attribute values are in English.

III.3 Structure and Functionality of TabML

28. TabML allows the storage of both document and data-related metadata. Document-level metadata serves to identify the document, relate it to other documents and place it in a wider context. Examples are the name of a survey, the reference period and creation date and time. Data-level metadata is specific to the data in a single document and describes the structure in which the data is embedded, like the number and type of columns and rows. However, the most prominent feature of TabML is that it comes with two different formats for tables, the Layout Format and the Matrix Format.

29. The Layout Format is designed for storing a table ready for print or display - in other words, for visualization - in a manner that enables its easy conversion into other presentational data formats like RTF, PDF and HTML, and it is therefore clearly regarded as a format for intermediate, transitory documents. It is similar to HTML, but richer in structure. It has more cell types and allows segmenting tables in order to describe and preserve page structures. The Layout format is also intended for storing the results of a layout generation from the Matrix Format.

30. The Matrix Format is considered the core of TabML. It offers a more general, abstract way of describing tables with a complete set of metadata, can hold any number of table matrices and relates each cell and its value to the metadata describing it. It is an all-purpose table format and not as restricted in its use as the Layout Format, but designed for supporting complex manipulations of the table such as statistical non-disclosure. However, it is rather intended for storage and information exchange than for direct processing.

III.4 Structure and Functionality of DatML

31. At the time being, the DatML development is still in an initial phase, and details of its structure cannot yet be given. Nevertheless, the following brings some of the functionality DatML will most probably have, though not all of it may be implemented in the first release.

32. In the first place, DatML will describe the structure of flat and hierarchical fixed and variable length data records, including names, single fields, composed structures and data types. DatML will achieve this through a relatively small number of basic element types that encapsulate the data.

33. Record structures consist of a small variety of basic elements, but vary extremely in how these elements are combined to form a record. DatML will offer ways of describing the constraints of a specific data set that allow to check if the basic data elements are arranged according to the record structure, and if hierarchical records appear in the proper sequence.

34. For the transmission of raw data and the support of data editing, DatML will provide elements that allow to store their results together with the data and the necessary document and data-level metadata.

35. For the purpose of data integration, DatML and TabML will share a common set of elements for document-related metadata.

IV. FUTURE OF THE STATISTICAL PRODUCTION PROCESS

IV.1 Consistent Metadata Administration with STATSPEZ

36. STATSPEZ will serve as a metadata administration tool at all stages of the statistical production. Its functionality and its metadata data base will be extended accordingly.

37. STATSPEZ will support the processing of paper questionnaires by supplying data entry forms and XML skeletons for statistical data.

38. The STATSPEZ metadata data base will be used to generate electronic questionnaires. It is still to be decided wether an existing XML document type such as IQML might be applied, but most probably, separate document types will be used for the questionnaire and - possibly DatML - the transmission of the raw data.

39. For data editing, STATSPEZ will allow the generation of program code in various languages (SPLV, Java, Blaise).

40. Appendix A provides a view on the future statistical production with STATSPEZ (see below).

IV.2 Application and Data Integration

41. The SPLV programs generated from the STATSPEZ metadata data base will both read and write DatML-conform data sets and output tables in both the TabML Layout and Matrix Format.

42. The GENESIS information system will use the TabML Matrix Format to supply data to and extract it from its data base. Tables intended for presentation will be output using the TabML Layout Format .

43. SAS will use DatML and the TabML Matrix Format for both input and output, and the TabML Layout Format for the presentation of tables.

44. The TabML Layout Format will be integrated into XML document types for publication.

45. With DatML and TabML having a common set of elements for document-level metadata, data integration becomes feasible for the major part of statistical data throughout the production process.

46. Document types for questionnaires, data entry forms, publications and others will be developed or existing ones will be resorted to. In either case, efforts will be made to include the same or a functionally equal set of elements to make these documents eligible for data integration with TabML and DatML documents.

V. CONCLUSIONS

47. The German Statistical Offices have powerful, metadata-enabled tools at their disposal that cover most of the statistical production. This allows to introduce XML document types for tables and statistical data - TabML and DatML – to further integrate these tools and take steps toward data integration.

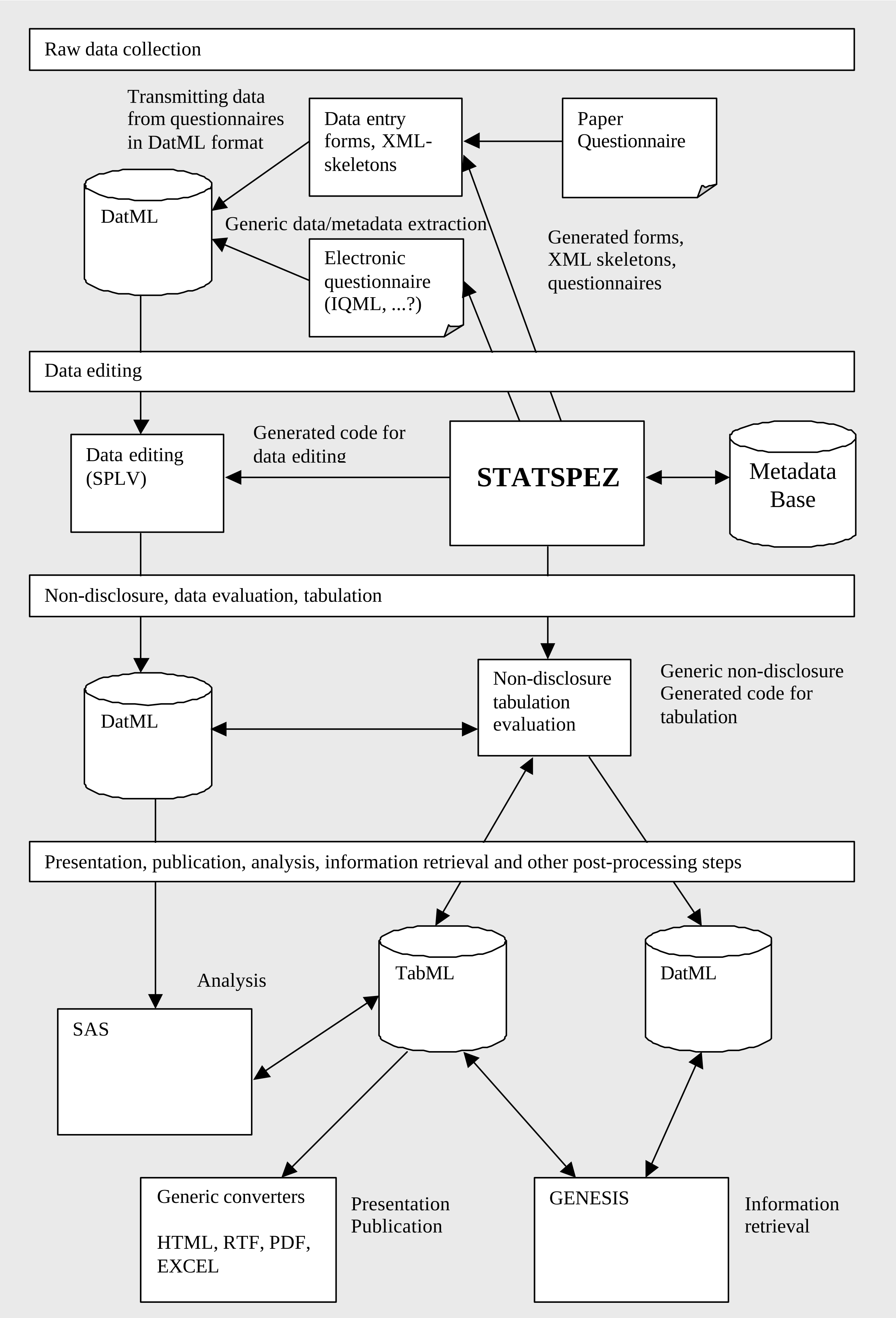
48. The existing STATSPEZ tabulation tool will become a unified tool for metadata administration. New functionalities will be added that extend the range of STATSPEZ onto all stages of the statistical production, and at the same time provide a means of administering and processing data and metadata consistently from data reporting to tabulation.

49. The XML-based document types TabML and DatML - and probably others - will be developed for data and application integration, and will provide a set of document types that allow to store metadata consistently.

50. TabML and DatML will be implemented into the GENESIS information system and third-party tools to make use of the metadata in the areas of information retrieval, statistical analysis and publication.

51. In the long run, the shift from application to data-centered processing will increase quality and efficiency of the statistical production and help make more of our data.

Appendix A: Future Statistical Production with STATSPEZ and XML


Raw data collection
Transmitting data from questionnaires in DatML format
Data entry forms, XML-skeletons
Paper Questionnaire
DatML
Generic data/metadata extraction
Electronic questionnaire (IQML, ...?)
Generated forms, XML skeletons, questionnaires
Data editing
Data editing (SPLV)
Generated code for data editing
STATSPEZ
Metadata Base
Non-disclosure, data evaluation, tabulation
DatML
Non-disclosure tabulation evaluation
Generic non-disclosure Generated code for tabulation
Presentation, publication, analysis, information retrieval and other post-processing steps
Analysis
TabML
DatML
SAS
Generic converters
HTML, RTF, PDF, EXCEL
Presentation Publication
GENESIS
Information retrieval