

Electronic statistical publication of the future

1) Changing over from a printed statistical system to an electronic one

Official statistics constitute a statistical system comprised of a collection of essential social statistics of high quality that are produced regularly and sufficiently frequently and are nationally representative¹.

General social statistics constitute the main product of Statistics Finland. At the moment Statistics Finland produces some 200 sets of statistics on 26 topics. Some of the statistics are compiled monthly or quarterly, some annually or less frequently. New statistical data are comprehensively released on the Internet. New data releases number almost 700 per year. The vast majority of the statistics produced by Statistics Finland are official statistics.

Statistical information has traditionally been reported, released and disseminated in the form of printed publications, which go back a long time. Statistics Finland's oldest statistical publication series have been produced for over 250 years. The evolution and development of the Internet has put the traditional way of reporting, publishing and disseminating statistical information in turmoil. The traditional form of publishing where tables and the information needed in their interpretation are bound into one volume in a publication series of official statistics is disappearing. NSOs are changing over to electronic dissemination of statistics, both as far as tables (databases) and the text analyses (publications) needed in their interpretation are concerned.

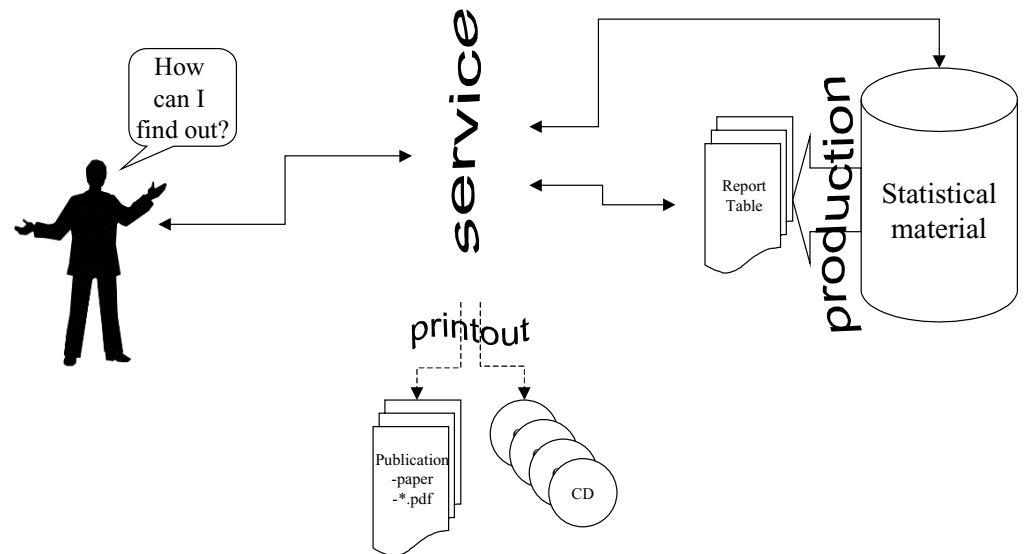
2) How to recognise a good electronic service

Thus, a major change of paradigm is taking place in statistics, away from dissemination of printed information to dissemination of electronic information. To quite a large extent the transition has already happened. In good decade, numerous innovative solutions facilitated by new technology have been created for the publication and dissemination of statistical information. New services have been developed to make up-to-date data available to users as fast as possible.

The changeover from paper products to electronic ones that technological progress has made possible has revealed problems in the old product-based concept. The limitation of the concept is that the user of statistical information must first find the product (publication or an electronic service) containing the information and the actual information can only be found after this. This represents to poor service. So that the possibilities offered by the Internet could be exploited and the conflicts arising from product-oriented thinking solved, NSOs must review their paradigms and change over from product-oriented to service-oriented thinking.

The essence in the service-based concept is immediate fulfilment of users' information needs. Implementation of the concept relies on a basic service operating as self-service, from where the user can exhaustively retrieve information with simple searches or queries.

¹ Laatusuhteissa (Quality Guidelines for Official Statistics). Käsikirjoja 43 (Handbooks 43). Statistics Finland 2007, p. 13.

Figure 1. Information search in the service-based model²

What kinds of characteristics should electronic dissemination of statistical information and the entire statistical system possess when the transition is made from printed to electronic dissemination? Statistical information must be exhaustively available in the service, it must be easy to find, published data must always be accompanied by the metadata needed in their interpretation and direct linking to them must be possible. In my paper I discuss problems, requirements and solution models relating to these issues.

3) Common Structure of Statistical Information (CoSSI)³

To implement the service-based model (Figure 1) the actual statistical information (tables, publications) and the metadata needed in its interpretation must be put into electronic format. However, this alone does not create a good service. Automatic conversion of information for different dissemination channels, its archiving, multiple language versions and attachment of sufficient metadata to it are all problems that must be solved in order to create a good service.

To solve these problems a Common Structure of Statistical Information (CoSSI) was developed. The point of departure in the CoSSI was an infological analysis of statistical information. 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. CoSSI defines the structures of statistical data, metadata and publications.

The CoSSI model is a modular DTD system. It consists of Document Type Definitions (DTDs), it is based on standards (CALS, XDF, Dublin Core), it has a DTD for statistical matrixes and a DTD for statistical tables, it has DTDs for publications and documents. One XML file contains data, metadata and all language versions.

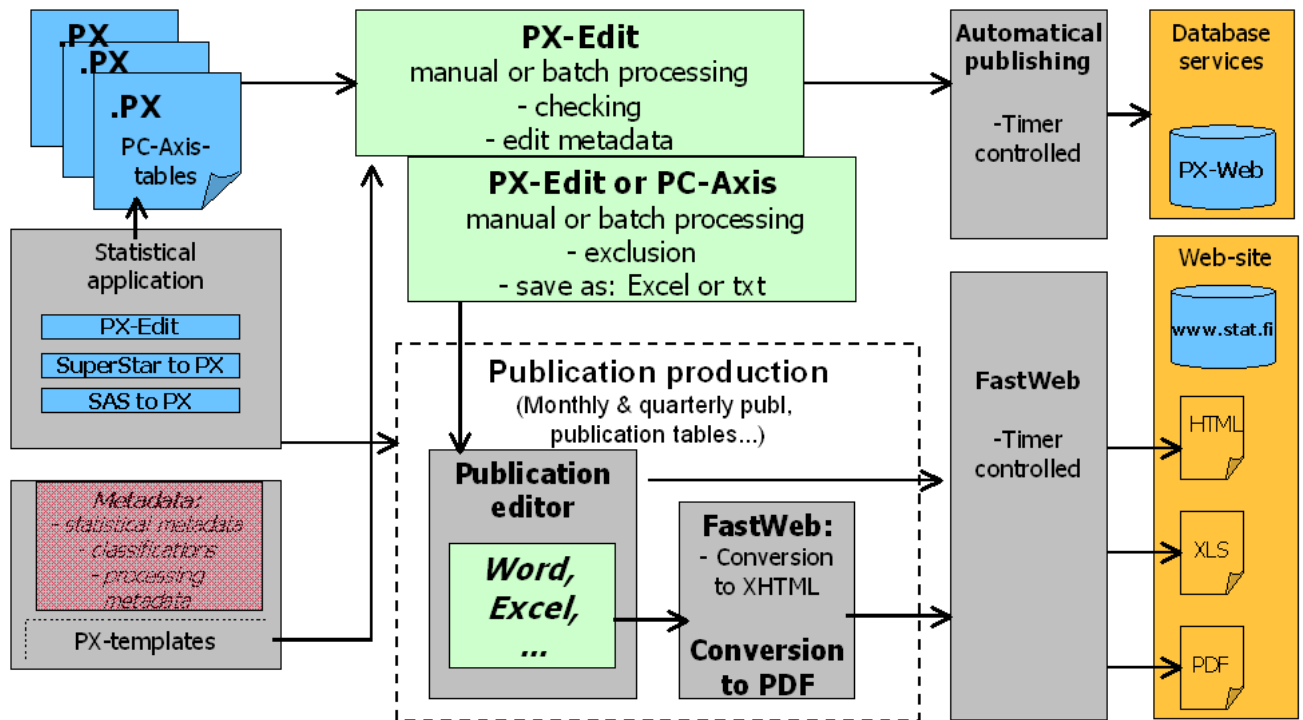
² Heikki Rouhuvirta, Markku Huttunen. How NSOs can respond to changing needs in the Internet era. Statistical Journal of the United Nations Economic Commission for Europe, Volume 20, Number 1 (2003), pp. 55-69, <http://ehournals.ebsco.com/direct.aso?ArticleID=NU7KV0Y64VL4FUUU29XP>

³ Heikki Rouhuvirta, Harri Lehtinen. Common Structure of Statistical Information (CoSSI) <<http://www.stat.fi/cossi>>.

4) XML publishing process compliant with the CoSSI model

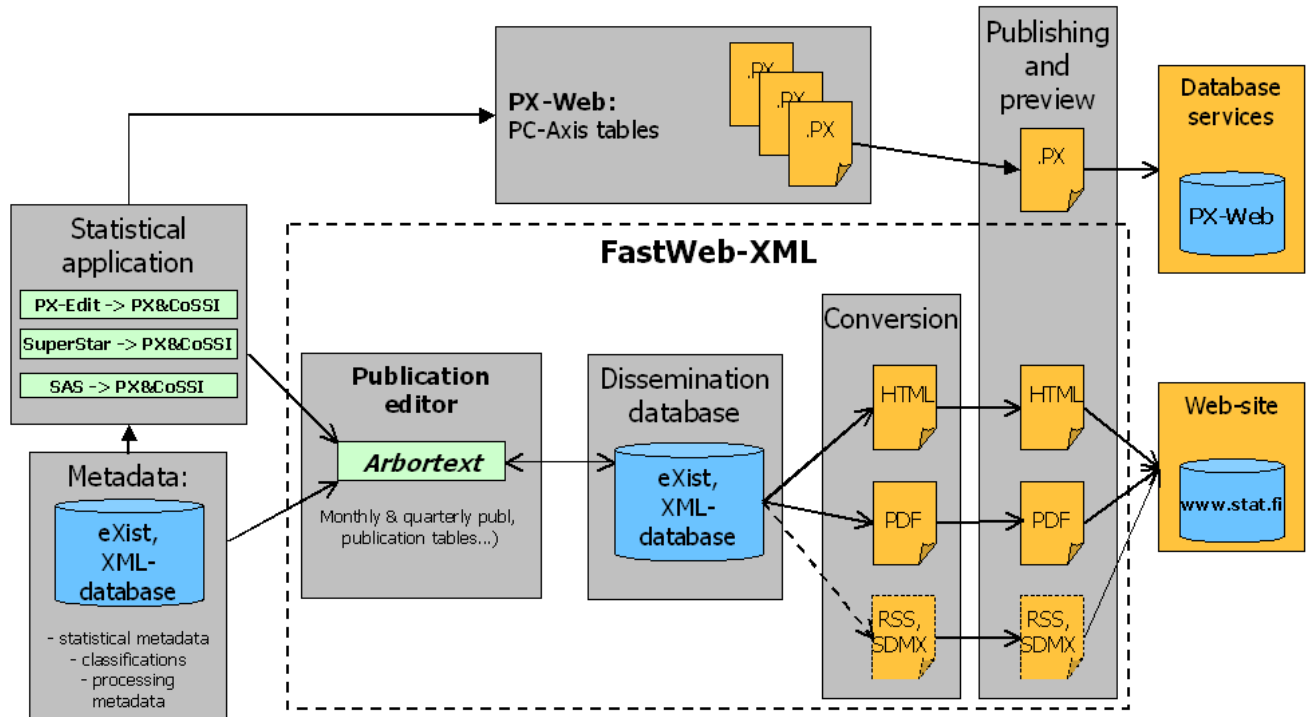
In Statistics Finland's old publishing process, publishing from databases, updating of www pages and production of publications (printed, PDF and HTML) were all their own, separate production processes. Different tools were used in each process and overlapping work was done in them (Figure 2).

Figure 2. Dissemination process – Office97



XML publishing process that complies with the CoSSI model is now at the introduction stage at Statistics Finland. In the new process, different formats of publications (printed, PDF, HTML) are generated automatically from one original XML file. Other forms of dissemination (email, RSS) are also produced automatically from the same original (Figure 3).

Figure 3. XML based dissemination process - XML and PC-Axis



At the first phase, database tables (.PX) are produced with the same production applications as the tables for the XML publishing process. SAS, SuperStar and PX-Edit produce both tables in XML format for the XML publishing process, and tables in .PX format for database dissemination purposes (PX-Web). At this phase the preview functions and publishing for different publication formats (printed, PDF and HTML), and database tables (.PX) are combined within one transmission application and the files are saved into one directory with a uniform structure. The same structure of topics and statistics becomes available in both the statistical www service and in the PX-Web database. The naming of XML, HTML, PDF and .PX files has been standardised at the different phases of the publishing process.

5) Archiving of an electronic statistical publication

Very little concern was given to the archiving of information published on the web when innovative services were being built in Internet's first decade of existence. The new situation has not been sufficiently examined by NSOs from the perspective of an entire statistical system in the long term either. How will the statistical information that is published today only in electronic format be available and accessible in, say, ten years' time? What about in 30 years' time? If we now needed Consumer Price Index data from 30 years ago, finding the data for 1976 would be possible with the help of a publication series of official statistics.

In the production process that conforms with the CoSSI model a publication original is saved in XML format in an eXist-XML database. This XML original of the publication contains the actual information content, the statistical metadata and the publication metadata in the desired languages in one XML file. This XML file is also archived into the XML database.

At the time of publication, intended dissemination formats (printed, PDF, HTML, etc.) are produced from this original XML file. The volume of metadata varies by publication format, for example, number of pages soon becomes a limiting factor in printed publications. The most extensive metadata content can be attached to a publication in HTML format, whereby the desired metadata are printed out auto-

matically as HTML files in addition to the HTML files containing the actual statistical data.

The different formats (PDF, HTML) of publications published in the www service are published so that the URL addresses generated for different files at the time of publishing do not change later on. This makes using the URL addresses elsewhere, such as diverse studies, articles or other web-based services, sensible and data user will always also be able to check the original document given in source references. Keeping published publications permanently accessible to users is an essential element of a good information service and system of official statistics especially if data are no longer published in printed form.

6) *Identification of electronic publication series*

Thanks to the Internet the printing of many statistical series has already been discontinued and publishing of the data has been transferred to the web. In addition, when new sets of statistics are added to production they are disseminated almost entirely electronically as tables in databases and publications in the www service.

Official statistics as a statistical system is seen in libraries around the world and in various sales systems through publication series identifiers (ISSN). Individual publications are often also identified with the ISBN publication identification system. These systems are world-wide standards and can be used for finding printed statistical publication series easily anywhere in the world.

As the numbers of printed statistical publications and series diminish, the statistical series (especially printed ones) identified in these systems give a deficient picture of the amount of statistical data produced by the system of official statistics. Examined through these systems the volumes of statistical series and published statistical information seem to be declining although in reality they keep growing continuously.

Both the ISSN series system and the ISBN identification system also facilitate identification of electronic publication series and publications. However, they have not been comprehensively introduced in electronic dissemination of statistics. The first reason is that these systems have not been adapted to also suit electronic publications until in recent years, and that their introduction demands from electronic publications very precisely defined properties and procedures in respect of, for example, making of corrections and supplementations. In addition, statistical data published in databases are totally outside the reach of these systems.

Electronic web documents can also be identified with the URN system, which can be combined with the ISSN series identification system and the ISBN publication identification system when necessary.

Under the current rules of the ISSN and ISBN systems, a statistical system and electronic data publishing and dissemination taking place within it can be brought under the scope of these conventional systems. By also introducing the URN system created specifically for the identification of web documents, even documents that do not necessarily meet the definition of a publication can be covered.

The use of all aforementioned ID systems requires that an electronic publication series and individual publications are unambiguously specified and their versions (e.g. corrected or supplemented ones) are unambiguously controlled. A miscellaneous collection of HTML pages whose contents change due to diverse updates, additions and deletions will neither be granted an ISSN series status nor an unambiguous publication ID (ISBN). Individual HTML documents may each receive an URN identification code but the actual collection of publications will not be specified.

7) Roles of an electronic statistical publication and table database

Along with statistical databases, the volume of statistical information published in tabular format has grown and continues to grow enormously. However, tables published in databases have a different role and different properties than tables published in printed or electronic publications. These differences deserve closer examining.

Tables published in databases are typically maintained as time series tables with continuously supplemented data contents. Database tables are also not archived in the same way as tables published in printed, HTML or PDF publications (or tables published in other file formats, such as Excel). For example, revisions of time series data, corrections of errors, or changes in classifications change the contents of a database table retrospectively even in respect of earlier data, sometimes as long as years after its first publication.

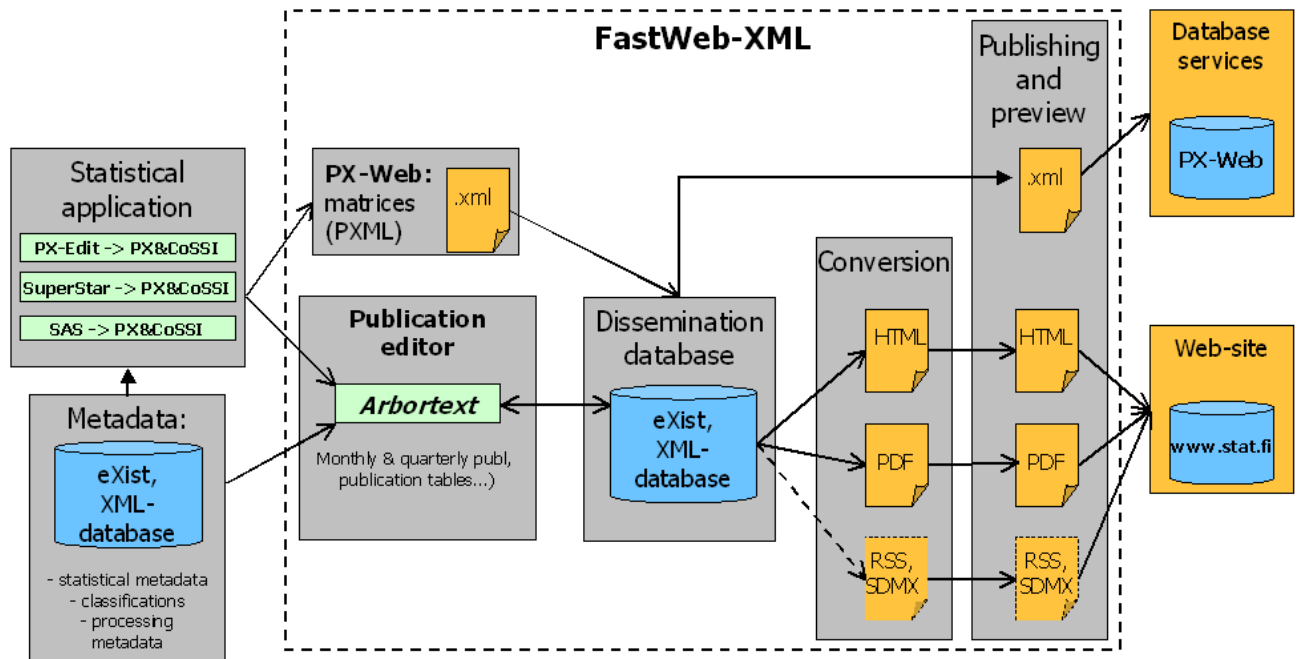
Thus, in database tables, data on even older observations are always published according to the latest situation. Because of this, time series published in databases may also deviate from data relating to older points in time which have been published as preliminary data in either a printed or an electronic publication or which have otherwise been revised or corrected subsequent to their first publication. For this reason, linking an electronic publication to the related database tables is not fully justifiable. The data only correspond with each other at the moment of publishing, for database tables may change later whereafter the data in them and in the electronic publication, and in the database no longer match.

At Statistics Finland, the tables created from the PX-Web database are not linked to any other background database, but the PX-Web database builds up on the server from published PC-Axis files instead. In connection with the introduction of the new XML publishing system each PC-Axis file published in the PX-Web server is automatically saved into a directory of statistics. Therefore, publications published on the home page of a set of statistics and the PC-Axis files saved into the same directory can easily be permanently linked with each other. Then PDF or HTML format publications published on the home pages of individual statistics and the related database tables (in PC-Axis format) become archived unchanged. This has two significant advantages: 1) it makes sense to also link the tables published in the database to the publication because they remain unchanged, 2) this way an archive is also created for the tables published in the PX-Web database.

8) Integration of an electronic statistical publication and a statistical database

Integration of the production processes of electronic publications compliant with the CoSSI model and database tables will become possible when the XML file format (PXML) is also adopted in PC-Axis and PX-Web programs (Figure 4).

Figure 4. XML-based dissemination process - integration completed



Then both the original XML files of publications and the original XML tables of database tables are saved into the eXist-XML database. Electronic publications can be easily linked to the related database tables when they are being produced, and both then archived as an interrelated entity into the eXist-XML database.

Electronic publications published on the website service are linked permanently to the related database tables, which are kept permanently accessible to users at their original, fixed URL addresses. Electronic publications are published under electronic ISSN series and individual publications are given ISBN and/or URN ID codes. When all database tables relating to an electronic publication are archived in the same way as the actual publication, the data published in the database are permanently accessible to users in the same form as at the time of publishing.

The storage and retrievability of data from a statistical system has thus been solved also in the long term. Electronic publications and database tables (and all data in them) published with them, linked with each other, and archived together with the publications can thus also be found in future through the ISSN, ISBN and URN systems. Thus, Consumer Price Index data for November 2007 can be found and accessed in, for example, 2037 exactly as they were when first published in December 2007.

Sources:

[1] Laatus tilastoissa (Quality Guidelines for Official Statistics). Käsikirjoja 43 (Handbooks 43). Statistics Finland 2007, p. 13.

[2] Heikki Rouhuvirta, Markku Huttunen. How NSOs can respond to changing user needs in the Internet era. Statistical Journal of the United Nations Economic Commission for Europe, Volume 20, Number 1 (2003), pp. 55-69, <<http://ejournals.ebsco.com/direct.asp?ArticleID=NU7KV0Y64VL4FUUU29XP>>.

[3] Heikki Rouhuvirta, Lehtinen Harri, Common Structure of Statistical Information (CoSSI) <<http://www.stat.fi/coSSI>>.