

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

Introducing an XML-based publishing process

Abstract

For a couple of years now Statistics Finland has been developing an XML-based publishing process and related tools for it. The production and publishing of XML-based publications has now been introduced in practice and statistical publications have been produced with the new system since the beginning of 2007. The transition to the new system is taking place gradually and the aim is that all statistics will be published by means of the new system by the end of 2008. This presentation describes the experience gained during development and introduction of the system from the perspective of the compilers of statistical publications. In addition, it gives examples of what the XML-based publishing process and related tools look like in practice.

Background

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. The 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 has 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 in which 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, as far as both tables (databases) and the text analyses (publications) needed in their interpretation are concerned.

For a long time, it has been the objective of Statistics Finland to publish more statistics on its web site, and the volume of data has indeed increased substantially every year. In 2000, Statistics Finland introduced a publication system, which converts Word documents to html pages. This system is still in use, although it was obvious from its introduction that it is not best-suited

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

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

for large-scale electronic and multi-channel publishing. Especially when used to produce large electronic publications in several languages, a large number of individual Word documents have to be created for conversion into individual html pages because of the system. It is very difficult to manage such projects and the workload is immense. For instance, electronic publishing of monthly statistics may require dozens of individual Word documents.

We obviously needed a more efficient content production system for wide-scale electronic publishing. On Statistics Finland's new web service all statistics it produces have their own, permanent home page and each statistic must also generate information for its page at least each time of publishing. We started to develop an XML-based publication system because we needed the capacity to produce large volumes of data, support for multi-channel publishing and dissemination in all required languages.

Structuring statistical data

Statistics Finland started to explore structuring statistical data in the early 1990s. The work was based on an existing and established SGML standard. A CALS table specification was used as the basis for the statistical tables. However, since the CALS specification was not sufficiently accurate to exhaustively structure statistical tables or material, features were added to it to describe the data content without disrupting the original specification. Work on the SGML was never completed, however. Statistics Finland was not fully prepared to accept new technology which could have had a wide-ranging effect on the systems it used to produce statistics, at the time.

Work was restarted in the late 1990s. SGML had already been succeeded by the XML standard and internet technologies were being rapidly developed. What is more, Statistics Finland was strongly expanding the volume of its electronic publishing, which introduced new requirements for publication systems and processes.

The work on SGML had been carried out with a very low profile and no particular effort had been made to commit the agency to it. A decision was made to restart the work and commission an external company to do a report which would then provide a foundation for a project to improve the dissemination of data produced by Statistics Finland. A Finnish consultancy specialising in XML technology and its potential in publishing statistical data was chosen for the report, which was completed in 2000.

XML pre-report

The consultancy was asked to study how XML could be used in disseminating statistical data. Statistics Finland's existing solutions for sharing information and the system it used to update its website, which was based on Office97, were the starting point for this work.

The consultancy's report analysed the status of XML and associated technologies, recommended that XML be adopted and suggested an XML

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

middleware solution where RDF would be used to create a consistent management system for documents produced at different sources. The solution would have created an RDF-based data description system, which would not, however, have influenced the actual publication production system, and hence would not have modernised it or corrected its defects especially concerning electronic publishing.

However, based on the report, we proposed to the management of Statistics Finland that an XML pilot project be launched, although it was given objectives that went beyond a mere XML/RDF-based solution. The aim was to test a purely XML-based solution in producing statistical publications to gather knowledge and experience, and justify the building of an XML-based publication system. The goal was also to prove that XML could provide a solution to produce a multi-channel publication from a single original in the formats required by different media.

XML pilot

The XML pilot was launched in 2001. A call for tenders was arranged for outside providers to help make a demo version of an XML-based publication system. The call was won by the Finnish company Republica.

Structural specifications

The basis of the pilot was a purely XML-based production of publications. We decided to look for structural specifications to specify statistical publications, tables and metadata. The SGML work done previously was the basis for this work. The SGML work provided the CALS specification for the tables and an XML version of it was adopted, with the extensions added as required for statistical data. The most interesting of the publication specifications was the Docbook definition. Docbook, however, proved to be too broad for statistical publications and we decided to make our own DTD specification. We did use the way Docbook organises the data content of publications and documents and in the end, the Statistics Finland specification for publication structure came out very Docbook-like. For metadata, the starting point was that metadata content must cover data content as specified by DublinCore. The associated data were entered in the metadata specification and special metadata elements were made for statistical publications and data.

The work was carried out so that Statistics Finland defined what data content was needed and the technical DTD writing was done by an outside supplier. DTDs for publications and metadata were written by Republica, while Citec, with whom we had collaborated in the SGML project, drew up the table specification. The outcome was a modular DTD system, the first version of the CoSSI model. At this stage, the model included the first publication.dtd version, statistical tables (expanded CALS model) and document metadata (docmeta.dtd.). From the first model onward, development of the DTDs was carried out solely by Statistics Finland.

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

From the onset, the structural specifications had to support multi-lingual publication, and they and the model had to suit all statistical information and be modular so that it would be simple to expand them with new data elements. Also from the onset, XML-DTDs were the technical platform.

Publication system based on the pilot project

Work on the structural specifications got off to a good start but building an XML-based publication system proved to be problematic, as did finding the right tools to produce XML documents. Both SGML and XML have been widely used for a long time in, for example, technical writing and technical manuals. The assumption was hence made that an XML-based publication system could be created on this basis.

- We did, however, encounter a few rather substantial differences and underestimated their effect. Firstly, technical manuals have a very strictly defined structure which is more narrow than the structural specifications we drew up for statistical publications, which covers all associated needs from small press releases to monthly and annual publications and statistical studies.
- Secondly, tables in technical manuals are so-called text or presentation tables and lack the content structure required of statistical data. This was observed with the CALS specifications, too, and it was the reason why we made the extensions. The problem was, however, that printing closely specified XML tables was not possible with the statistics production system and adding a printing function to different systems proved to be more difficult than anticipated.
- A third significant difference is that technical manuals are written by technical writers whose main occupation is writing. In other words, they use the tools and editors needed to produce XML documents on a daily basis and master even the more complicated ones. At Statistics Finland, however, statistical publications are produced by people whose main occupation is something else and they need tools that are easy and quick to embrace. Despite the many XML editors currently available, it was very difficult to find a suitable one.

The principal achievement of the XML pilot project was that it produced the first structural specifications. The solution to structure statistical data that was the basis of the structural specifications and the modular DTD solution chosen for structuring are still in use, although they have been expanded considerably.

The first versions of the XSLT and XSL-FO conversion scripts were another significant result from the pilot project. They work with documents that comply with the structural specifications. Thanks to the specifications and the scripts, we were able to show that with the XML technique, structured statistical data can be disseminated across multiple channels so that different versions of a single original (publication, press release, table) can be produced for different channels and information needs. In the pilot project,

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

XML was converted into HTML pages and PDF files and a demo version was made of an e-mail transmission to deliver the lead of a publication to customers. In addition, the project implemented an application named X-Fetch which handles the publication and conversion process. X-Fetch was used to determine what conversion would be made to XML documents, where the source documents would be sourced and where they would be saved, and to comprehensively manage the XML-based publication process and associated applications.

As I mentioned above, we experienced difficulties with the publication editor during the pilot. The original hope was that Microsoft Word (Office97) could be tailored into an editor suitable for producing XML documents. At the time Statistics Finland was using – and still is – a publication system based on Office97 in which documents are structured using Word styles and converted to XHTML. The idea was that this solution would be expanded to produce XML documents that would comply with the structural specifications of Statistics Finland. Documents cannot be saved in XML directly in Word but solutions do exist as Word add-ons, which make this possible. The project also tested Up-Cast software, which uses styles to convert Word documents to XML as defined by the user. The problem was, however, that Word does not steer or force users to apply styles correctly, and as a result, it was very difficult in practice to produce a Word document that could be simply converted into a valid XML document. The same problem also applies to new versions of Word when the aim is to produce an XML format that is not the one internal to Word. Moreover, the pilot project lacked the resources to modify an existing XML editor to suit statistical publication and hence the issue was not resolved.

Results of the pilot project

The pilot project proved that statistical information can be structured into XML and produced the first version of a general Statistics Finland structural specification for statistical data. It was also proven in practice that XML format documents are very well suited as the foundation of a multi-channel distribution system.

However, because of the problems with the publishing editor and the XML printing of the statistical production systems, the findings of the pilot project could not as such be applied to production. A brief time-out was therefore taken after the project before work on the publication system was continued in the Statistics Finland Production Model project.

XML for statistical data

Work on XML structural specifications of statistical material was carried out alongside the pilot project and then later the production model project. A CALS specification exists for statistical and publication tables but it is very ineffective in saving large statistical material in XML format, for example. To improve the model, it was expanded with a new module (matrix.dtd),

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

which allowed much XML material to be saved without expanding the file size beyond manageable.

This structural specification, which is also known as the matrix form, is based on a NASA-developed XDF.dtd, which we modified to better suit statistical material. The result was a structural specification with which the statistical data and their metadata could be converted into XML. Its information content was a full match with our expanded CALS specification with its metadata but technically, it was very similar to, for example, the PC-Axis file format.

With the matrix specification, we were able to show that XML could be introduced deeper in statistics production. We also created scripts for conversions between the CALS table format, the matrix format and PC-Axis files. We also created a demo version of an editor to view XML matrices and CALS tables.

Project to modernise publication production

A project to modernise publication production was started at the end of 2002 under the Statistics Finland Production Model project. Statistics Finland needed a replacement for its Office97-based publication production system. Also, Statistics Finland had increased and was continuously increasing the statistical information available on its web site and the old system was no longer well-suited for producing data content. The findings of the pilot project were sufficient to convince Statistics Finland of the usefulness of XML and that it would be worthwhile investing in its further development.

The project to renew publication production continued where the XML pilot left off and also made use of other experience gathered along the way on XML techniques and tools.

The starting point of the second project was that the structural specifications drawn up in the pilot would define the structure of the documents worked on. The specifications were naturally expanded as the work progressed by introducing new data elements but the basic solution has remained the same to this day. Hence the conversion scripts with which XML documents are modified for various channels could also be used.

Otherwise, most of the experience gained from the pilot indicated which paths were not worth following. It did not pay to continue looking for some kind of middleware solution where the data content would first be produced in one format and then converted to XML. As a result, we abandoned publication editor solutions based on Office97. We also tested newer Office versions, which do have limited XML support, but they did not satisfy our needs, either. This meant that the only remaining option was to find an editor that would natively produce XML and work in compliance with our structural specifications and be sufficiently easy to use for people who make statistical publications.

In addition, we gave up on the X-Fetch application. It proved to be very laborious to administrate and adopt and adding new functions to it was very

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

challenging. Moreover, we began to suspect that all functionality provided by X-Fetch could be provided by much simpler and scalable solutions.

The actual dissemination database was not dealt with in the pilot project, although the Tamino XML database was tested, but it was never linked to the pilot environment in any way. However, we planned to look for an XML database or some traditional database with features of the XML database.

Tables in XML format

In the pilot, we did not deal with XML printing from statistical applications. We should have, however, because not having done the work caused considerable delays to the testing of the system we were developing. We took up the issue in the production model project and succeeded in printing XML publication tables that complied with the structure specifications using SAS, SuperStar and Pc-Axis/PX-Edit/PX-Web. However, this was not really within the scope of the publication modernisation project. Responsibility had not really been assigned to anyone and hence managing the task was problematic. Teams formed to carry out the work that were outside the project organisation itself and consisted of people interested in it. This caused problems and delays, however, because sufficient resources had not been allocated for the work and it was done in addition to other responsibilities.

Currently, XML publication tables can be printed from PX-Edit, PC-Axis and PX-Web and also from SAS and SuperStar. XML printing for PX-Edit was done by Statistics Finland and for Pc-Axis/PX-Web by Statistics Sweden. Hence the CoSSI XML table specifications also became the official XML format for the PC-Axis family. Statistics Finland also tailored the XML printing function of SAS, known as *ODS - Output Delivery System*, to produce XML tables, and now all tables made with the *proc tabulate* procedure are directly available in XML. As for SuperStar, Space Time Research created an XML table printing functionality based on our specifications.

This means that key statistical applications used by Statistics Finland can currently produce the table needed for publication in the required XML format. However, there is still work to do. At the moment, we can produce tables in one language but it would help publication to have all language-versions in the same table. It would also help if we could produce the metadata describing a table's content with the table. Work is in progress on both of these issues and we expect to have results later in 2007.

Publication editor

We needed an editor that could natively produce XML, use our structure specifications and be easy to use for a person who works with statistics, is not familiar with XML and doesn't make publications very often. This proved to be rather difficult. We organised a call for tenders based on which we chose the Arbortext editor, which used to be named Epic. Arbortext did

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

not suit our needs out of the box and it was tailored for us, as was promised in their tender.

To implement Arbortext means that it must be tailored to comply with the structure specifications and usage. The editor offers a selection of applications and others can be tailored directly by programming. The tailoring was given to an outside supplier and Statistics Finland defined the needs and required functionalities, and the supplier provided them. The following were included in the editor:

- Styles for publications, tables and metadata
- Database connection to the XML database
- Automated functions to aid the writing of publications
- Support for multilingual documents
- Separate windows for adding and editing metadata

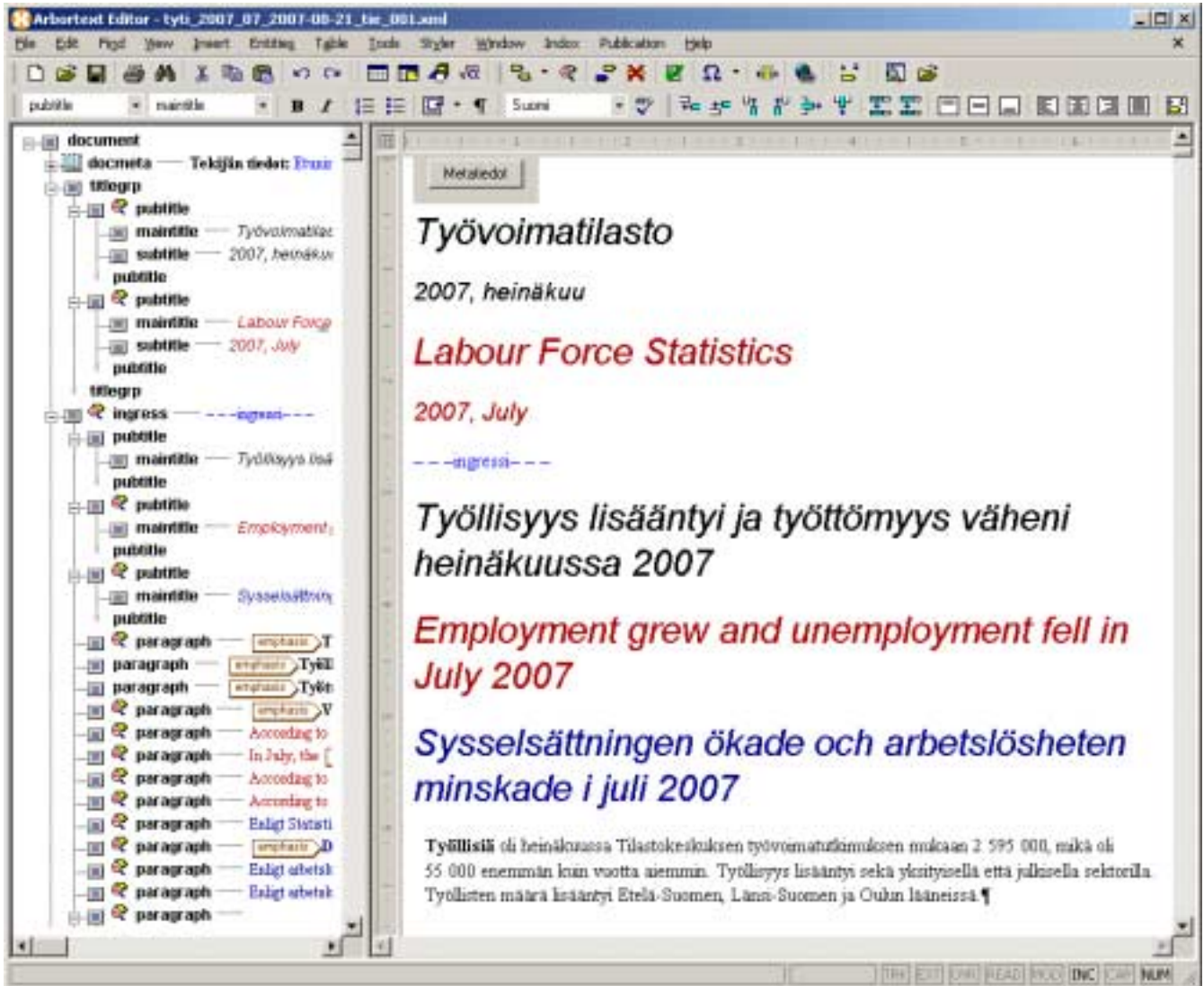
We collaborated closely with the tailoring supplier and learned to do tailoring ourselves, too, which was important because our plan is that Statistics Finland staff will do all future development work.

Arbortext ensures that documents produced with it comply with structural specifications. Its continuous validation function also includes an automated function that assists writers by providing them with real-time information on which structural elements may be added to a particular place in a document. Writers do not need advanced competence in XML syntax and structural specification in a technical sense. Only general knowledge of the content of the structural specification is needed.

We also want to make the interface as similar to Microsoft Word as possible to make it easy to adopt. The styles were specified to correspond, as closely as possible, to the styles Statistics Finland has specified for Word. The principal difference with Word is, however, that while Word allows its users to apply styles rather freely, Arbortext does not. We seem to have done rather well with the publication editor, and with the first courses behind us it now looks like a one-day course is enough to provide users with sufficient skills to make publications with the new editor and in a new environment.

12.9.2007

Information Technology and Statistical Methods
 Harri Lehtinen
 harri.lehtinen@stat.fi



Dissemination database

The pilot project tested the Tamino XML database. Tamino is a commercial product, however, and includes a sizeable number of features we felt were not needed in the publication production system. Hence the pilot project used network drives to pilot XML-based publication production. We knew, however, that this would be a temporary solution for the needs of the pilot only.

In the project to renew the publication production system we took a broader look at database solutions for the management of XML documents. The alternatives were conventional relation databases with an XML extension, a commercial XML database or an open source XML database.

At the time, conventional databases did not have very developed XML support, especially the ones in use at Statistics Finland, and we decided to abandon this alternative. The remaining alternatives were native XML databases, which we compared with each other. The comparison included the main commercial and open-source products. For us, the key difference between commercial and open-source versions, aside from price, was

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

version management, which is much better in the former than in the latter. We decided on an open-source product in the end because we did not want to invest funds at the time and did not consider version management to be that much of an issue. We chose the eXist database.

eXist is easy to implement and modify. Currently, it is used as a data store and an archive for XML-based publication production, and we are in the process of making it a data store for XML metadata and data gathering. We added direct eXist access to Arbortext to allow users to load documents directly from the database to the editor and then save them back in the database.

In addition, eXist feeds the documents forward in the publication chain. Every time a publication document is saved, the database releases a trigger that guides the XML document through the changes (XSLT & Saxon, XSL-FO & XEP) to be previewed in HTML and PDF format. The content of the eXist database is also backed-up in real time and can be used as an electronic archive for all XML-format documents.

Because the eXist database has very good support for XML inquiry languages, document searches and indexing is effective and easy. What is more, we have not noticed any performance issues associated with the eXist. We tested the database before it was implemented by adding a large part of the Statistics Finland HTML site into it and experienced no problems with searching and indexing this mass of some 15,000 HTML documents.

Currently, the only significant defect we have discovered is that there is no version management. The eXist database can be tailored and modified, however, and we intend to add at least a simple management system using the techniques available in the database. Of course, we could change to a commercial product with version management in the future.

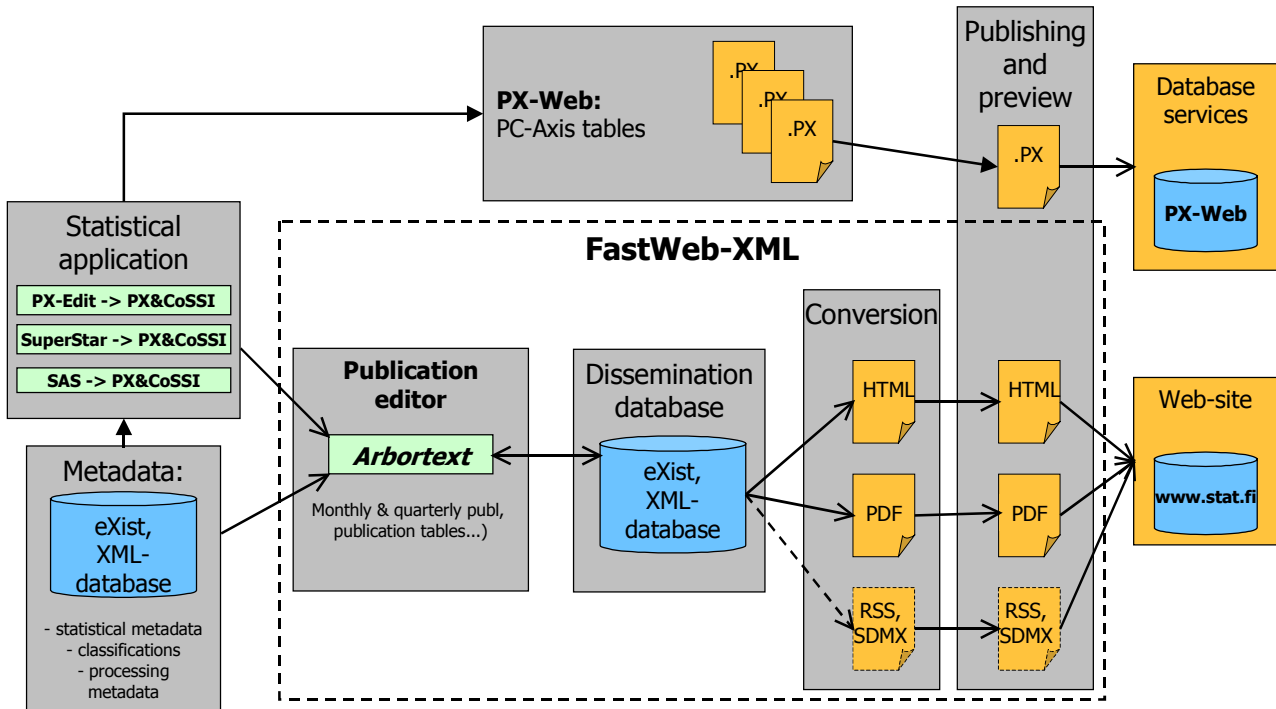
Publication production process

The process of publication production took the following model in the development work.

12.9.2007

Information Technology and Statistical Methods
 Harri Lehtinen
 harri.lehtinen@stat.fi

/ XML based dissemination process – XML and PC-Axis



The objective of the process is a workflow that is as simple, manageable and clear as possible from creating statistical tables to the final publication of the information through various publication channels. The final shape of the project was the result of the general model of the statistics production process on the one hand and the tools and their features that were added during the development work, as well as the solutions they made possible.

The publication production process is based on publication tables, which are always in the same technical format (XML tables) irrespective of the application used for the statistics. This is the key interface between producing and disseminating statistics. Thanks to the consistent table format, we could build a dissemination system where it was not necessary to modify applications used to produce statistics in connection with the dissemination modernisation, with the exception of XML output.

The key component in producing publications is the publication editor (Arbortext), which is used to write them and save them in the database (eXist). The editor is used to add tables and diagrams in the publication, fill in associated metadata and write or compile the texts.

The eXist database is used as a data store that feeds the publications forward in the process to conversion applications and to the final publishable formats. The final application in the process is the preview and timing interface that is used to preview the publication in its final formats (HTML and PDF) and set the final time of publication.

12.9.2007

Information Technology and Statistical Methods
 Harri Lehtinen
 harri.lehtinen@stat.fi

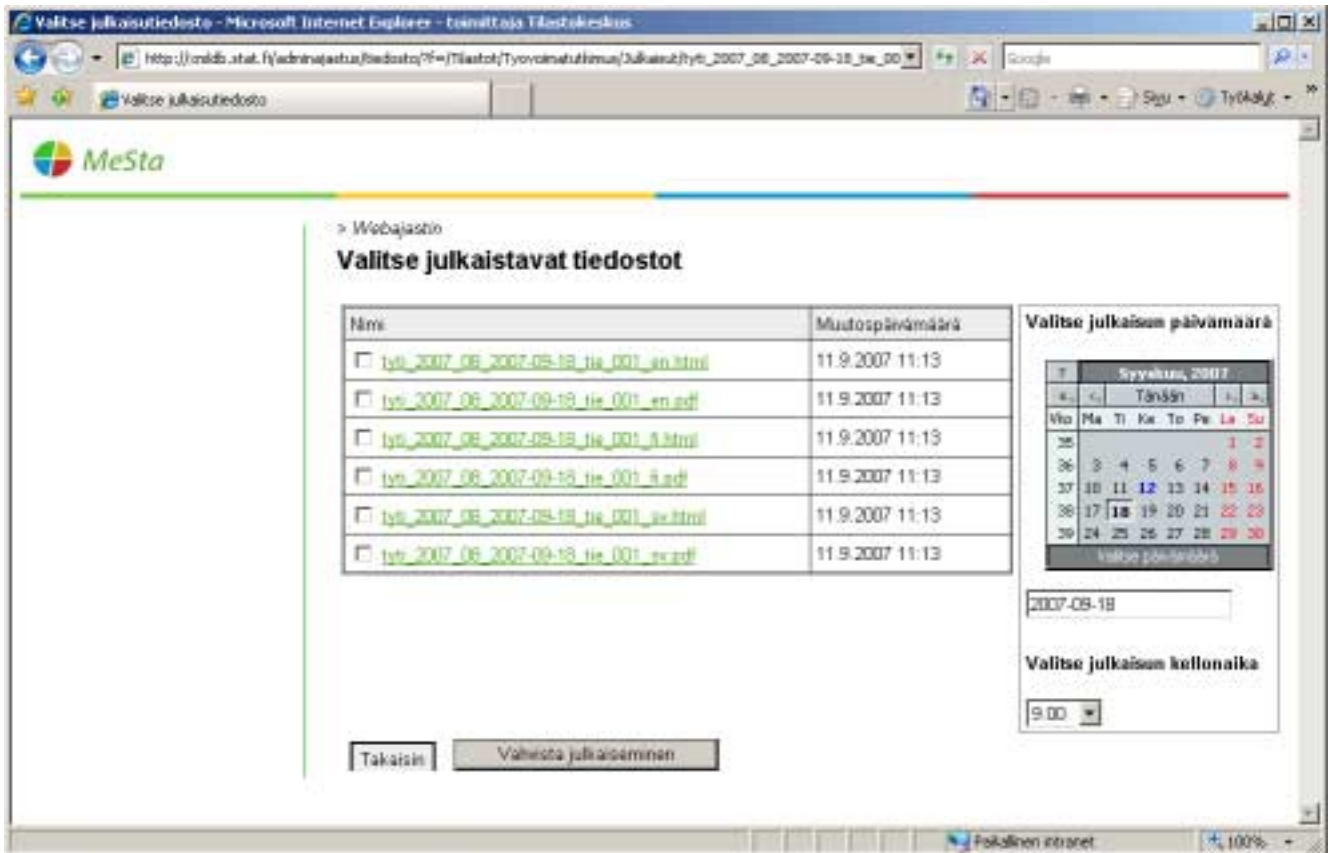


Illustration 3. Preview and timing interface

Implementation

The users of the new publication system and those who make publications work in the various statistics departments. In other words, publications are still produced in several departments, as was the case under the previous publication system. On the other hand, while some 200 people had access to the old Office97 system at Statistics Finland, the number of people up-dating the system should be lower under the new one and production will be concentrated on fewer people.

The XML-based system was introduced in production in March 2007. By the end of September 2007, a total of 32 monthly, quarterly and annual statistics from 18 statistics had been published and the aim is that all statistical publication will take place under the new system by the end of 2008. Statistics published under the new system by the end of September include:

- Labour force survey
 (http://tilastokeskus.fi/til/tyti/2007/07/tyti_2007_07_2007-08-21_tie_001_en.html);
- Enterprise openings and closures
 (http://tilastokeskus.fi/til/aly/2007/01/aly_2007_01_2007-07-26_tie_001_fi.html);

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

- Telecommunications
(http://tilastokeskus.fi/til/tvie/2006/tvie_2006_2007-06-05_tie_001_en.html);
- Adult education survey
(http://tilastokeskus.fi/til/aku/2006/aku_2006_2007-05-21_tie_001_fi.html);
- New orders in manufacturing
(http://tilastokeskus.fi/til/teul/2007/07/teul_2007_07_2007-09-10_tie_001_en.html);
- Preliminary population statistics
(http://tilastokeskus.fi/til/vamuu/2007/07/vamuu_2007_07_2007-08-16_tie_001_fi.html);
- Statistics on income and property
(http://tilastokeskus.fi/til/tvt/2005/tvt_2005_2007-03-30_tie_001.html);
- Producer price indices for services
(http://tilastokeskus.fi/til/pthi/2007/02/pthi_2007_02_2007-07-17_tie_001_en.html); and
- Statistics on central government productivity
(http://tilastokeskus.fi/til/vatt/2006/vatt_2006_2007-06-21_tie_001_fi.html).

Before starting to work with the new system, publication producers will be given a one-day course. They will also receive personal induction when producing their first publication with the new system, if they need it. This will include a step-by-step walk-through of publication production, with special emphasis on the relevant statistics production system. This has been quite sufficient in most cases, and participants have been able to make their second publication independently.

Because the XML-based publication system is used to produce both web pages and a PDF file from the same original, tables published with them will have to be re-thought. For instance, the size and form of tables for monthly publications were previously dictated by fact that they were printed on paper. Now the same table will have to work on the web, too. When we transferred the labour statistics publication to the new system, for example, we redesigned every table in it and are now able to automatically produce the web site, the printable PDF version of the entire data for the site and another PDF file for the printing house from a single original.

Conclusion

We are now using an XML-based process to produce publications (FastWeb-XML) which has proven itself in practice. However, we are still working to improve the diagram production and metadata in particular. For the diagrams, we have a temporary solution in place which produces reasonably good quality for publication, but there is still room to improve the process and tools. Statistics Finland will launch a project to this end later in 2007. The metadata work has been carried out in a metadata project whose aim is to convert statistical metadata to XML format. This will allow

12.9.2007

Information Technology and Statistical Methods
Harri Lehtinen
harri.lehtinen@stat.fi

us to integrate and automate metadata more closely with publication production and eventually also with statistical production.

According to the Statistics Finland work-hour logs, the development process has required the following amount of work hours:

Year	Work hours
2007	2019
2006	3252
2005	4402
2004	2267
2003	1364
2002	213
2001	1097
Total hours	14614
Total days	2088
Total months	104
Total years	9

This table does not include all the other work in support of systems development because it has not been entered accordingly. An estimated 30 people have taken part in the work at different stages and in different projects. Some of them have worked full-time while others have invested fewer hours in the development of a particular component, for example. Programming has also been commissioned from outside suppliers especially to tailor the publication editor.

Development work will continue especially with regard to statistical graphics and metadata. We are also fine-tuning the interface of the editor on the basis of user feedback.

The biggest hurdles in implementation have been learning to use the editor, and adapting the data content and especially the statistical tables to the new system. However, this is not only affected by the publication system but also by the new website of Statistics Finland and the strategic decision to make electronic publishing the principal channel. In fact, the principal function of the XML publication system is to support the transition to electronic publishing.