

Multichannel publishing of official statistics – a challenging task

Abstract

Statistics Finland's new XML publication process is now used in production and the basic publications from statistics are being transferred to the new process. Many new procedures will be applied during the changeover. The changes are not just technical, but e.g. the contents of publications need to be redesigned as well. While especially the designing of tables in the electronic publication system creates new demands, it also offers new possibilities. The producers of publications enter data into structured publication files, whose contents will be distributed in many different ways via various electronic channels. What is required of the producers of publications?

In connection with the revision of the publication procedure the series division of official statistics will also be renewed; printed publication series will be replaced with electronic publications and tables published in the table database. The relationship between electronic publications and the table database brings its own challenges e.g. as regards the archiving of data. The users of the Internet require data as quickly as possible and in an easily digestible form, but at the same time they should be able to interpret and use the published data correctly. What demands does this place on the producers of the online service?

In this paper, I will examine the requirements of adopting an XML publishing system and the effects it will have on the content of publications, their distribution and archiving, the process of producing publications from the producer's perspective, the functionality of the online service and the series division of official statistics. The technical documentation and the XML DTDs of the XML publishing process can be found from Statistics Finland's online service.¹

1. Background

Statistics Finland renewed its practices for releasing statistics in Finnish in 2004, and in Swedish and English in 2005². According to the new publishing practices, a standard format release is produced for all statistics whenever new statistics are compiled and the data are made available to users. Before 2004, new statistics were published in press releases which numbered some 300 every year. The new publishing practices have made the publishing of data more comprehensive. About 650–680 statistical releases have been produced annually, which shows the considerable increase in the number of releases.

The new procedure was adopted for Swedish and English releases in June 2005. With the changeover, the number of Swedish language statistical releases grew from the earlier approximately 120 press releases to the same

¹ Common Structure of Statistical Information (CoSSI); <http://www.stat.fi/rossi>

² Statistics Finland's new English web pages released (26 July 2005);

http://tilastokeskus.fi/ajk/poimintoja/2005-07-26_webpages_en.html

International Marketing and Output Database Conference
1 September - 5 September, 2008, Naantali
Markku Huttunen (markku.huttunen@stat.fi)

number as in Finnish, i.e. about 650–680 statistical releases, from the start of 2006. The number of English language statistical releases has increased from about 120 press releases annually to 330 statistical releases in 2008³. The objective is to increase the number of statistics released in English further.

The modernisation of the publishing practices has taken into account the objectives and requirements of the XML publishing process currently under construction from the perspectives of e.g. archiving and multichannel distribution. All statistical releases are archived in the online service and their URL addresses will remain unchanged.⁴ The structure of statistical releases was carefully defined to meet the demands of multichannel distribution. For example, the headings of statistical releases and the first paragraph are automatically delivered to the e-mail and RSS distributions.⁵

Statistics Finland's XML-based publishing system became operational in the spring of 2007. The monthly publication of the Labour Force Survey, which served as the pilot in the XML project, has been created with an Arbortext XML editor since 22 May 2007 and published on Statistics Finland's online service in HTML and PDF format in both Finnish and English (with only the statistical release in Swedish), as well as in printed format in Finnish.⁶ A single three-language XML original file can be used to automatically publish HTML pages in Finnish, English and Swedish (a total of about 90 HTML files), compile a PDF publication in Finnish and English, compile a PDF file in Finnish for printing and distribute the publication via e-mail and RSS in Finnish, English and Swedish.⁷

By July 2007, the publications of 35 of the good 200 Statistics Finland statistics were being created in both HTML and PDF format with the new XML process. So far the new publishing process has been used to produce 120 different publication titles in Finnish, English and Swedish (in both HTML and PDF format).

Redesigning of publications

In the old publishing process, the data contained in a single set of statistics was published 1) as electronic statistical release in the online service (the compulsory minimum; if required, annexed tables and figures and a longer article would also be published), 2) as a printed publication and 3) as database tables in the table database. All three publishing processes were separate and employed different tools in the production of tables and the publishing of data. Data in different distribution channels were not co-ordinated and they did not form a consistent whole.

The text section of the old printed Labour Force Survey was a two-column publication in Finnish. The tables were bilingual (Finnish, English) and were

³ Release Calendar 2008;

http://tilastokeskus.fi/ajk/julkistamiskalenteri/julkistamiskalenteri_aika2008_en.html

⁴ Statistical release archive; http://tilastokeskus.fi/til/arkisto/index_en.html

⁵ Latest statistical releases from Statistics Finland; http://tilastokeskus.fi/media/rss/2.0/tk_en.rss

⁶ Content of the Labour Force Survey's home page is expanding;

http://tilastokeskus.fi/til/tyti/tyti_2007-05-22_uut_001_en.html

⁷ Labour Force Survey > 2008 > June; http://tilastokeskus.fi/til/tyti/2008/06/tyti_2008_06_2008-07-22_tie_001_en.html

optimised for A4-sized pages. The publication was available for a charge in both printed and PDF format. Alongside the printed publication, a statistical release and some annexed tables and figures produced in a different publication process and containing different data were released free of charge in the statistics online service on home page of the statistics.⁸ The revised Labour Force Survey is a comprehensive publication composed of the text section, an extensive table and figure annex and a quality description.⁹ The publication is available free of charge in both HTML and PDF format, and a printed version of it can be ordered for a charge.

The new publishing system will combine the content of statistical releases published in the online service (and the table and figure annexes supplementing them) and the content of earlier printed publications. This revision is not merely a question of routine copying of earlier data into new tools, but a complete redesigning of the content as a whole. The objective is to retain the publication form in the packaging of published data also in the future.

Basic publication from statistics

The role of publications being transferred over to the new publishing system is to release basic data on the official statistics. A basic publication is an entity that describes a single set of statistics at a certain reference time and it includes the statistical release, a more extensive text section, table and figure annexes and a standardised quality description. If necessary, several basic publications from a set of statistics can be released at a certain reference time, including a preliminary release and releases containing final data or, for example, several publications focusing on different topics.¹⁰

The objective is to transfer the basic publications from statistics comprehensively into the new XML-based publishing process. The majority of statistics will be converted to the new system between 2007 and 2009. In addition to technical changes (incl. the modernisation of table production so that they are produced in the XML format), the content of publications will be redesigned as part of the changeover (texts, tables, general structure of the publication), the agency's publishing policy will be revised (charges for publications, whether or not a publication will be available in printed format, etc.) and the series division of official statistics will be rethought.

Structured publications

The new publishing tool (Arbortext XML editor) will be used to compile a publication into a single structured XML file. The publication's language versions will be edited in the same editor window, and all language versions will be saved in the single structured XML file. This XML file will be used to automatically publish the different content combinations of the publications through different publishing channels on the release date. The producer

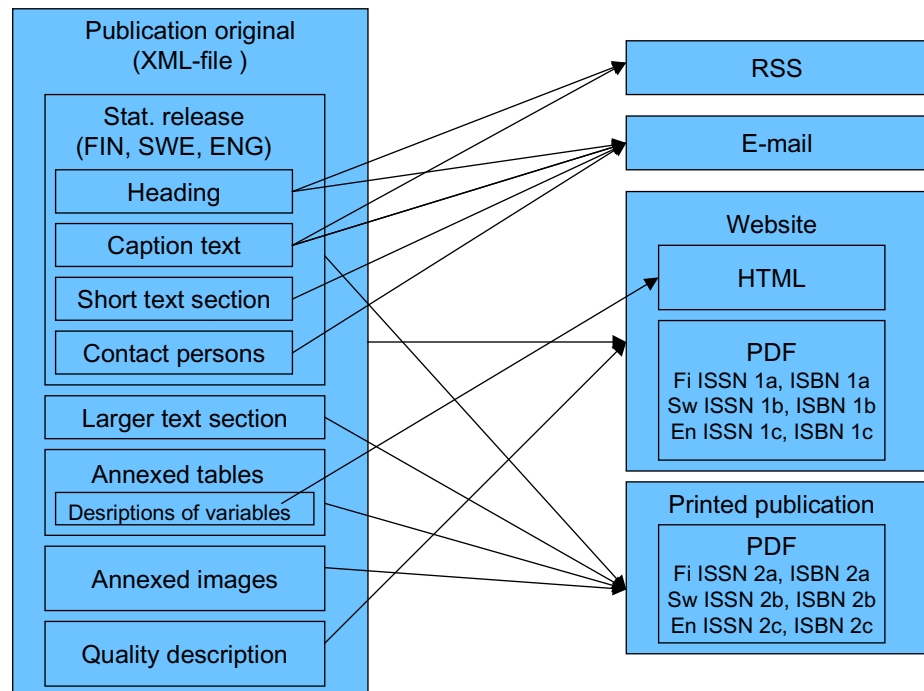
⁸ Labour Force Survey 2007, March (statistical release in all languages, appendix tables only in Finnish); <http://tilastokeskus.fi/til/tyti/2007/03/index.html>

⁹ Labour Force Survey 2008, June; http://tilastokeskus.fi/til/tyti/2008/06/tyti_2008_06_2008-07-22_tie_001_en.html

¹⁰ Adult Education Survey 2006 (only in Finnish and Swedish); http://tilastokeskus.fi/til/aku/2006/01/aku_2006_01_2008-06-03_tie_001_sv.html

of the publication will not need to worry about the distribution of the content when drawing up a publication; it will be sufficient that he/she compiles the publication in accordance with the agreed structure. Language versions will be processed using the structural components of the publication, i.e. the content of a publication's different language versions will be managed at the element level.

Figure 1. Single XML original file -> several content combinations for different distribution channels in different languages



Simultaneously and automatically, a single XML file -> HTML, PDF, printed publication, e-mail, RSS, different language versions

Redesigning tables

The role of tables in a basic publication from a set of statistic is to describe the key points in a compact way. Even though PDF versions of publications are also created for multichannel distribution, the principle is that the electronic version is to be created first and foremost. The data content of a publication is optimised primarily for effortless use in an Internet browser. A publication's tables, therefore, must function properly in HTML format, which requires that they be small and deal with only one subject.¹¹

Although the rationale behind small tables is based on the obvious space and usability restrictions, the use of small tables can be justified also on the grounds of information service. The message from usability experts is clear: The users of Internet services want to access information on a website quickly and in an easily understandable format. Large and complicated tables do not meet such information needs. The role of figures is also to depict the key phenomena in a compact way.

¹¹ Telecommunications, Telecommunications 2007; http://tilastokeskus.fi/til/tvie/2007/index_en.html

International Marketing and Output Database Conference
1 September - 5 September, 2008, Naantali
Markku Huttunen (markku.huttunen@stat.fi)

such a way that the links remain unbroken when the publication and its database tables are archived.¹² This will also enable the creation of permanent links between the compact publication tables and the large source tables published in the database as well as their archived versions.

Revision of the series division of official statistics

The series division of Official Statistics of Finland used to be based on the printed publication series. With publications going increasingly electronic due to advances in the Internet, no new printed publication series have been established in recent years. In addition, old printed publication series have been discontinued at an ever-increasing rate. Therefore the series division of official statistics has not been comprehensive enough to describe the developments taking place in all areas of society. Moreover, the constant shift to electronic publishing is reducing coverage all the time.

The Official Statistics of Finland were redefined and are now based on the division of statistics. The data contained in a set of statistics can be published in various forms (printed, electronic, in a database, etc.) as well as in various publication series (printed publication series, electronic publication series). As each set of statistics is transferred to the new electronic publishing method, it will be essential for maintaining the continuity of the series division of official statistics that the new electronic publications are also published in the official statistics series. Hence, the data can be found using the old series division of official statistics (SVT-ISSN) in the future as well.

OSF-ISSN (Official Statistics of Finland) serial numbers for Statistics Finland's statistics will be obtained as each set of statistics is transferred to the new publishing system. This will mean that the data contained in each set of statistics will be published under a specific section in the official statistics series. When the changeover becomes comprehensive, the new statistic-by-statistic series division of official statistics will comprise almost 200 series. Each language version will be given a separate serial number, so if the content of the electronic publishing system becomes comprehensive in all three languages, there will be a total of almost 600 series in Finnish, Swedish and English.

Archiving of electronic publications and publication series

The long-term storage of printed publications for access by users has been intuitively easy to plan and carry out. Libraries have stored the publications they have ordered in accordance with their own practices. In the last instance, the National Library of Finland and the Library of Statistics archive and store all available printed publications and give users access to them. Even though gaining access to old publications may require some effort, the printed publications are available to everybody for research purposes, for example.

¹² Population by gender and area 31.12.2007 and increase of population (archived PC-Axis file); http://www.stat.fi/til/vaerak/2007/vaerak_2007_2008-03-28_tau_101_en.px

International Marketing and Output Database Conference
1 September - 5 September, 2008, Naantali
Markku Huttunen (markku.huttunen@stat.fi)

Once printed publication series have been discontinued and electronic publishing has been completely adopted, it will be necessary to archive and store electronic publications for long-term access by users. The availability of new publications has improved compared to earlier ones, as more and more publications are available online free of charge. But what will the situation be like after a long period of time, say 30 years? Or what if one wishes to find an electronic publication which is only five years old and which has been referred to in an article or publication, for the purpose of verifying the original source? After all, this is clearly one of the ways of using official statistics, and it is obvious that this should be possible.

From the user's point of view, the best service would enable him or her to find an electronic publication even years later – for checking an original source, for example – at its original URL address. If the URL is changed the publication may still be found in the publisher's online service using a search or indexing function. This, however, becomes increasingly difficult as time passes from the publication. If a publication in, for example, PDF format has been published in an ISSN series of a set of official statistics and an ISBN or a URN code has been obtained for it, the publication can be found and identified even if its original URL is no longer functioning. In future, publications with such ID numbers will be available at the National Library. The National Library archives publications for research use, and is not responsible for maintaining general and easy online access to them.

The situation becomes more difficult when data have been published only in HTML format and the publication does not have an ID number. If the structure of the online service has changed and the original URL addresses are broken, it will be extremely difficult or even impossible to check e.g. the references made to the publication later on, even if the original publication or parts thereof could be found.

As stated earlier, the adoption of electronic publishing will also mean that some of the table material, which was earlier published in print, will be transferred to a database. With the use of databases, the amount of data published has grown considerably due to the removal of the space restrictions of printed publications, but there have hardly been any solutions to the problem of long-term storage and findability of database publications.

From the point of view of archiving, what would the data contained in an electronically published set of official statistics have to be like in order for the service to be good? The following criteria, for example, should be fulfilled:

1. each publication should remain available at its original URL address (best findability),
2. publications should be marked with unambiguous IDs (ISBN, URN) (facilitates finding them even if the URL has changed),
3. if a publication has been corrected, the corrections should be unambiguously marked in the original publication so that the original data can be traced; or if a new corrected version of a publication has been created, both the incorrect original and the corrected new version should be made available and each should have a reference to the other so that us-

International Marketing and Output Database Conference
1 September - 5 September, 2008, Naantali
Markku Huttunen (markku.huttunen@stat.fi)

- ers could access all the necessary information concerning the corrections (the creation of different versions for corrections in electronic format is challenging once the time span becomes long enough),
4. tables published in a database should also be archived at their original URL addresses and they should follow all the requirements set out in the point above, including the fact that each version of the database table (revisions, corrections of errors) should also be stored and available for viewing,
 5. the data contained in long-ago discontinued publication series should also be available (incl. tables published in a database), which means the accumulating history of publication series/statistics should be stored for access.

Statistics Finland has striven to meet these requirements in the data architecture of its statistics website with the following means, among others:

1. all statistics publications and material related to them are archived automatically at their original URL addresses, for example: Consumer Price Index; http://tilastokeskus.fi/til/khi/tie_en.html
2. the publications released under the new system will be published in the ISSN series of official statistics and marked with an ISBN (annual publications), for example: Telecommunications 2007; http://tilastokeskus.fi/til/tvie/2007/tvie_2007_2008-06-05_en.pdf
3. for example Statistics on local government productivity, 2005, first published 31 January 2007 and corrected 12 October 2007; http://tilastokeskus.fi/til/kktu/2005/kktu_2005_2007-10-15_tie_001_en.html
4. for example Families by type in 1950-2007 (Excel table), revised 2 June 2008; http://tilastokeskus.fi/til/perh/2007/perh_2007_2008-05-30_tau_002_en.xls
5. the electronic publication series of discontinued statistics and the data contained in them will be stored for access by users on the online service after they have been discontinued, for example: Harmonised Index of Consumer Prices; http://tilastokeskus.fi/til/ykhi/tie_en.html

The archiving of electronic publications is not only a question of storing individual publications but also the management and archiving of the entire lifespan of a publication series on the online service. Between 2004 and 2008, Statistics Finland published 212 statistics on its online service, of which 204 were in production in August 2008. New statistics are launched, old ones discontinued, and statistics are combined or separated. What will be the total number of statistics compiled by the year 2020, for example, and how many of them will still be in production? The published and archived data contained in all discontinued statistics must be available for users in the future as well. Early in 2008, Statistics Finland combined three very frequently used statistics: the Harmonised Index of Consumer Prices¹³, the Cost-of-Living Index¹⁴ and the Consumer Price Index. The two former ones were merged into the statistics on the Consumer Price Index, in whose series

¹³ Harmonised Index of Consumer Prices; http://tilastokeskus.fi/til/ykhi/index_en.html

¹⁴ Cost-of-Living index; http://tilastokeskus.fi/til/eki/index_en.html

International Marketing and Output Database Conference
1 September - 5 September, 2008, Naantali
Markku Huttunen (markku.huttunen@stat.fi)

the data are now being published.¹⁵ The data of the discontinued statistics can still be found and used through the archived statistics series.

Interpreting data – metadata

Metadata – variables, concepts and definitions, classifications, quality descriptions, etc. – are essential to the correct interpretation and utilisation of data. In the past, the publication of metadata was restricted in printed publications by the amount of space available. The metadata published as part of a printed publication used to be easily available and remained so in archives as well. Various more extensive metadata publications that served as a user's handbook were also easy to archive. The space available for publishing metadata documents is no longer a restriction in the electronic format; all relevant metadata documents can now be published. Electronic metadata documents should be equipped with the necessary IDs in the same way as the publications themselves in order to make the information easier to find.

A simple way of ensuring that the metadata required for the interpretation of data always accompanies the data itself is to publish both in the same package. This also applies to electronic publications. For this reason, among others, publications produced using Statistics Finland's XML publishing system shall always have a quality description attached to them. Matching each quality description version to the publication itself is easy when they have been combined in a single package.¹⁶ The list of concepts and definitions is published through the concept database and corresponds to the status of the latest publication; the archive is not available to the users of the information.¹⁷ By attaching the concepts and definitions section to each publication it would be easy to ensure that the appropriate metadata always accompanies the publication. On the other hand, classifications are often so extensive that it is not practical to attach them to a publication in their entirety. In such a case it is important that earlier classification versions are stored for access by users.¹⁸ These should then naturally be also published in such a way that the entire classification can be stored at its original URL addresses and made available to users in the future as well.

¹⁵ The Cost-of-Living Index and the Harmonised Consumer Price Index will be published in connection with the Consumer Price Index as from January 2008;
http://tilastokeskus.fi/til/khi/khi_2007-12-13_uut_001_en.html

¹⁶ Migration, 2007; http://tilastokeskus.fi/til/muutl/2007/muutl_2007_2008-05-23_en.pdf

¹⁷ Migration, Concepts and Definitions; http://tilastokeskus.fi/til/muutl/kas_en.html

¹⁸ Industrial Classification - versions; http://tilastokeskus.fi/meta/luokitukset/toimiala/versio_en.html