Seminar on dissemination of statistics and launching a new web
solution in NBS, Moldova
Chisinau, 27. May 2010
Markku Huttunen (markku.huttunen@stat.fi)

Multichannel publishing of statistics (electronic publications and database) - Finnish experience

1. The consept of multichannel publishing

1.1. 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 representative.

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 about 650 per year.

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 evolvement 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.

What kinds of characteristics should electronic dissemination of statistical information and the entire statistical system posses 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.

1.2. Common Structure of Statistical Information (CoSSI)[1]

To implement good Internet-service the actual statistical information (tables, publications) and the metadata needed in its interpretation must be put into electronic format. However, this alone is 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

[1] Heikki Rouhuvirta, Harri Lehtinen. Common Structure of Statistical Information (CoSSI); http://www.stat.fi/cossi

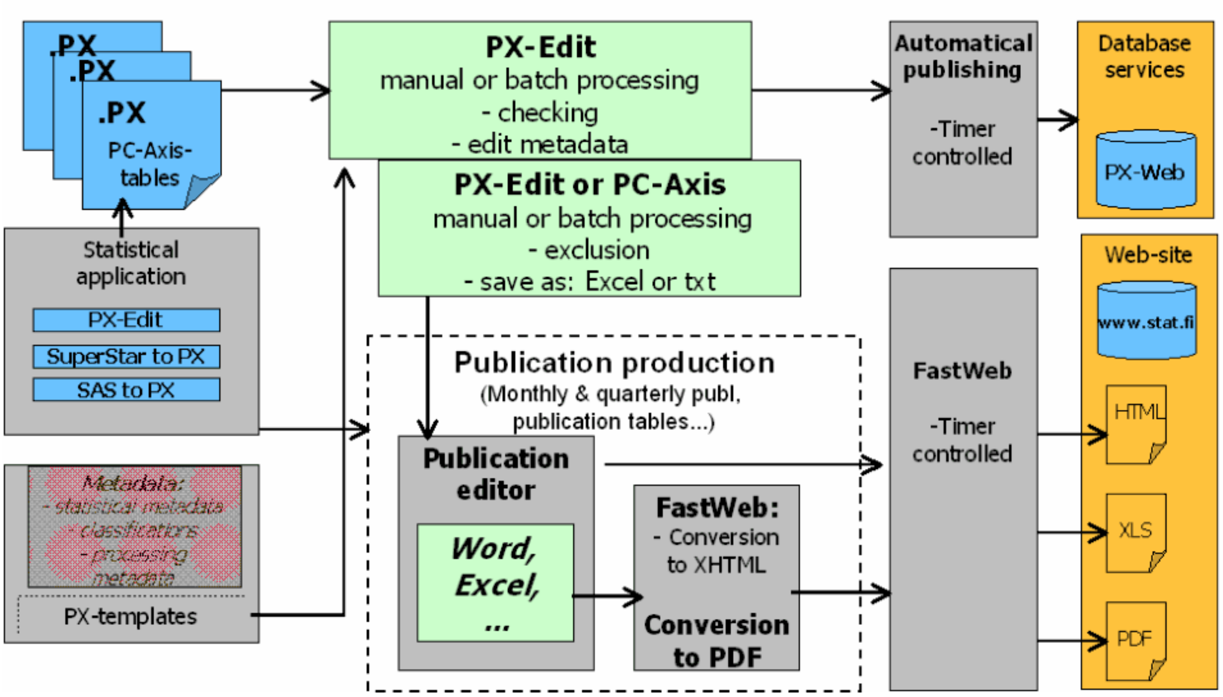
situation and application, in reality statistical information has a certain simplifiable and acceptable universal structure. The CoSSI describes the general structure that is not dependent on the situation of the statistical information presented in differing formats. CoSSI defines the structures of 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.

1.3. 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 1).

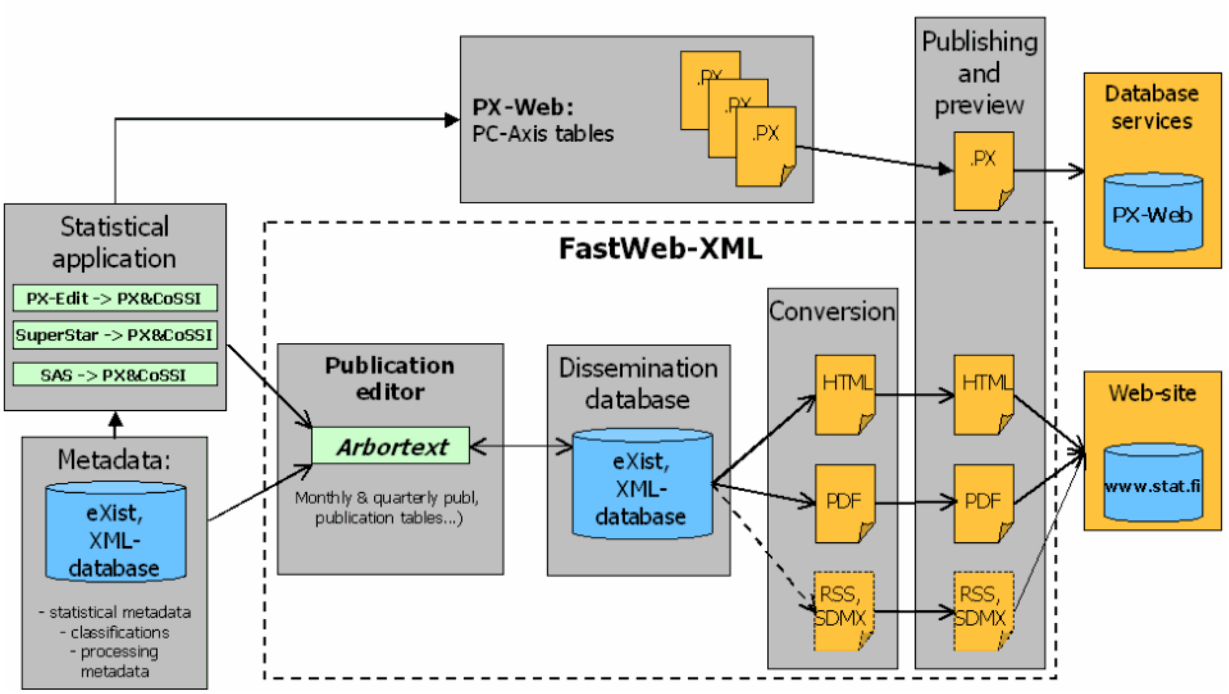
Figure 1. Dissemination process – Office97[2]



XML publishing process that complies with the CoSSI model will be fully implemented in the dissemination of 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 2.).

[2] Still used in 40 statistics in May 2010; example "Consumer Price Index"; http://www.stat.fi/til/khi/index_en.html

Figure 2. XML based dissemination process - XML and PC-Axis[3]



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.

1.4. 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 1980 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

[3] Used in 150 statistics in May 2010; example "Prices of Dwellings"; http://www.stat.fi/til/ashi/index_en.html

printed publications. The most extensive metadata content can be attached to a publication in HTML format, whereby the desired metadata are printed out automatically 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.[4]

1.5. 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 long 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.

[4] Statistical release archive; http://www.stat.fi/til/arkisto/index_en.html .

2. Multichannel publishing in practice in Statistics Finland

2.1. New practices for releasing statistics

Statistics Finland renewed its practices for releasing statistics in Finnish in 2004, and in Swedish and English in 2005[5]. 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 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 350 statistical releases in 2010[6]. 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.[7] 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.[8]

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.[9] 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.[10]

By May 2010, the publications of 150 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 over 1000 different publication titles in Finnish, English and Swedish (in both HTML and PDF format).

[5] Statistics Finland's new English web pages released (26 July 2005); http://tilastokeskus.fi/ajk/poimintoja/2005-07-26_webpages_en.html

[6] Release Calendar 2010; http://tilastokeskus.fi/ajk/julkistamiskalenteri/julkistamiskalenteri_aika2010_en.html

[7] Statistical release archive; http://tilastokeskus.fi/til/arkisto/index_en.html

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

[9] Content of the Labour Force Survey's home page is expanding; http://tilastokeskus.fi/til/tyti/tyti_2007-05-22_uut_001_en.html

[10] Labour Force Survey > 2010 > March; http://tilastokeskus.fi/til/tyti/2010/03/tyti_2010_03_2010-04-27_tie_001_en.html

2.2. Redesigning of publications

In the old publishing process (Figure 1), 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 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.[11] The revised Labour Force Survey is a comprehensive publication composed of the text section, an extensive table and figure annex and a quality description.[12] 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 (Figure 2) 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.

2.3. 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[13] or, for example, several publications focusing on different topics.

The objective is to transfer the basic publications from statistics comprehensively into the new XML-based publishing process. All 200 statistics in production will be converted to the new system by the end of 2010. In addition to technical changes (incl. the modernisation of table production so that they are produced in the XML format), the content of publications are redesigned as part of the changeover (texts, tables, general structure of the publication), the agency's publishing policy has been revised (charges for publications, whether or not a publication will be available in printed format, etc.) and the series division of official statistics has been rethought.

2.4. 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 cre-

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

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

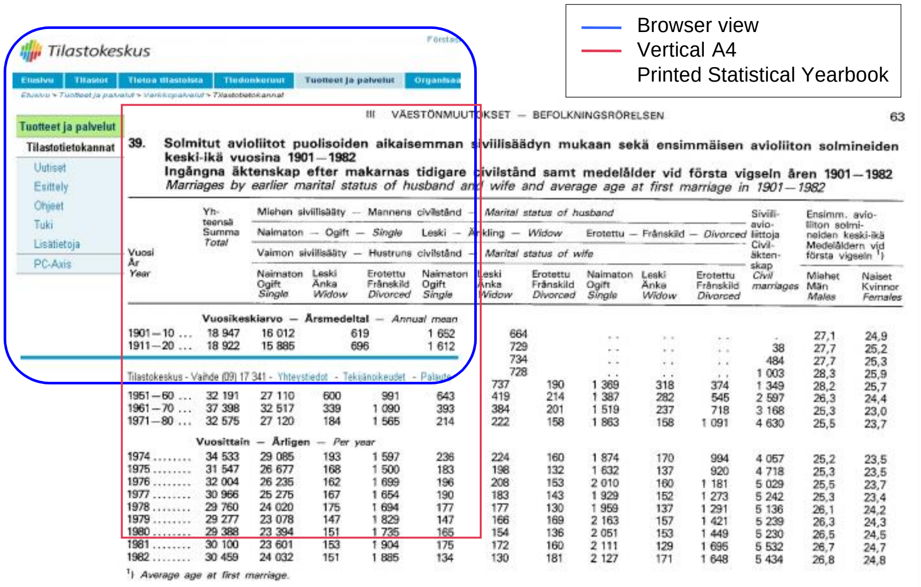
[13] Financial Accounts; http://tilastokeskus.fi/til/rtp/tie_en.html

ated 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.[14]

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.

It has always been convenient to create figures to represent a single aspect, so there is no need for major changes on their part when publications are redesigned. In any case, the number of figures should be increased considerably, because their properties make them well-suited for online use. The lack of advanced graphics creation tools has been an obstacle to increasing their number. Such tools, however, are being sought at the moment.

Figure 3. Three forms of a table: printed, A4, browser



III VÄESTÖNMUUTOKSET — BEFOLKNINGSRÖRELSEN 63

39. Solmitut avioliitot puolisoiden aikaisemman siviilisäädyn mukaan sekä ensimmäisen avioliiton solmineiden keski-ikä vuosina 1901—1982
Ingångna äktenskap efter makarnas tidigare civilstånd samt medelålder vid första vigseln åren 1901—1982
Marriages by earlier marital status of husband and wife and average age at first marriage in 1901—1982

Vuosi År *Year*	Yhteensä Summa *Total*	Miehen siviilisääty — Mannens civilstånd — *Marital status of husband*									Siviiliavioliittoja Civiläktenskap *Civil marriages*	Ensimm. avioliiton solmineiden keski-ikä Medelåldern vid första vigseln ¹)	
		Naimaton — Ogift — *Single*			Leski — Änkling — *Widow*			Erotettu — Frånskild — *Divorced*					
		Vaimon siviilisääty — Hustruns civilstånd — *Marital status of wife*											
		Naimaton Ogift *Single*	Leski Änka *Widow*	Erotettu Frånskild *Divorced*	Naimaton Ogift *Single*	Leski Änka *Widow*	Erotettu Frånskild *Divorced*	Naimaton Ogift *Single*	Leski Änka *Widow*	Erotettu Frånskild *Divorced*		Miehet Män *Males*	Naiset Kvinnor *Females*
	Vuosikeskiarvo — Årsmedeltal — *Annual mean*												
1901—10 ...	18 947	16 012	619		1 652	664		. .	. .	. .	-	27,1	24,9
1911—20 ...	18 922	15 885	696		1 612	729		. .	. .	. .	38	27,7	25,2
						734		. .	. .	. .	484	27,7	25,3
						728		. .	. .	. .	1 003	28,3	25,9
						737	190	1 369	318	374	1 349	28,2	25,7
1951—60 ...	32 191	27 110	600	991	643	419	214	1 387	282	545	2 597	26,3	24,4
1961—70 ...	37 398	32 517	339	1 090	393	384	201	1 519	237	718	3 168	25,3	23,0
1971—80 ...	32 575	27 120	184	1 565	214	222	158	1 863	158	1 091	4 630	25,5	23,7
	Vuosittain — Årligen — *Per year*												
1974	34 533	29 085	193	1 597	236	224	160	1 874	170	994	4 057	25,2	23,5
1975	31 547	26 677	168	1 500	183	198	132	1 632	137	920	4 718	25,3	23,5
1976	32 004	26 235	162	1 699	196	208	153	2 010	160	1 181	5 029	25,5	23,7
1977	30 966	25 275	167	1 654	190	183	143	1 929	152	1 273	5 242	25,3	23,4
1978	29 760	24 020	175	1 694	177	177	130	1 959	137	1 291	5 136	26,1	24,2
1979	29 277	23 078	147	1 829	147	166	169	2 163	157	1 421	5 239	26,3	24,3
1980	29 388	23 394	151	1 735	165	154	136	2 051	153	1 449	5 230	26,5	24,5
1981	30 100	23 601	153	1 904	175	172	160	2 111	129	1 695	5 532	26,7	24,7
1982	30 459	24 032	151	1 885	134	130	181	2 127	171	1 648	5 434	26,8	24,8

¹) *Average age at first marriage.*

- Figure: Jaakko Laakso

The role of database tables is significant when switching over from old -style printed publications to electronic publishing. Tables with extensive data contents must either be dismantled into small (browsable) tables dealing with one aspect or transferred to the database for publishing. In practice the majority of the tables in publications, which are larger than a single A4 -sized page, will be transferred to the database.

Therefore the basic publication from a set of statistics can be divided into electronic content published online, with two different formats and user interfaces (HTML and PDF), and tables published in a statistics database. The producer of a publication faces a considerable challenge in compiling the publication in such a way that it functions properly in all of its formats and in different operating situations. This is a case of multichannel distribution in its full scale.

The availability of official statistics over time is greatly reduced if tables published in the database are not archived and stored in a place where users can access them

[14] Telecommunications 2008; http://www.stat.fi/til/tvie/2008/index_en.html

for a long period of time. For its table database, Statistics Finland uses the PX-Web database application, which does not have any links to any background databases; a PX-Web database is based on static PC-Axis files published on a server. Statistics Finland has started the archiving of different versions of database tables published in PC-Axis format, which will enable a publication representing a set of statistics at a certain time and the database tables related to it to be permanently interlinked in such a way that the links remain unbroken when the publication and its database tables are archived.[15] 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.

2.5. 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.

3. Sources:

[1] 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.

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

[15] 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
[16] Migration, 2009; http://tilastokeskus.fi/til/muutl/2009/muutl_2009_2010-04-22_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