European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May to 3 June 2016

Training for compilers of statistical releases

Kim Huuhko[1], Erja Seppänen[2]

[1] Statistics Finland, Helsinki, Finland; kim.huuhko@stat.fi
[2] Statistics Finland, Helsinki, Finland; erja.seppanen@stat.fi

Abstract
At Statistics Finland, statistical experts produce statistical releases themselves and publish them online. There are around 160 active statistics which produced around 550 releases in year 2015. During 2014 and 2015, Statistics Finland organised an extensive training round for statistical experts that covered all statistics. The authors of statistical releases were invited to the training sessions as small groups and examples were used to review the recommendations for compiling statistics.

Keywords: Statistical releases, Presentation of data, Standardised metadata

1. Statistical releases and databases roles for conveying statistical data

1.1. Background for the training sessions

Many sub-areas related to Statistics Finland’s statistical releases have changed in recent years so new instructions and thorough training sessions were needed. Old instructions were several years old and thus, in some cases, were already badly outdated. New instructions and the reasons for changing them were explained in the training. At the same time, as all the instructions were revised they were also transferred to Statistics Finland’s new intranet, which was taken into use early in 2015.

Also Statistic Finland’s web site’s new graphic design was introduced in early 2015. That created some relevant changes for the release recommendations. For example, a narrower text column created tighter restrictions for the size of the individual html tables and graphs. As a part of the new graphic design Statistics Finland’s new colour palette was also introduced.

SAS is a strategic software most commonly used at Statistics Finland for tabulating micro data into aggregated dissemination tables for both the PX-Web databases and for the individual html tables of the statistical releases. SAS is also a widely used software for creating statistical graphs for the releases. New ways to create these graphs have been introduced into SAS in recent years. The other most commonly used software to create graphs for releases is Microsoft Excel. A new version of Excel was taken into use in late 2014. Changes in these software along with changes to the colour scales generated a need to revise all the colour palettes and graph templates for both software.

Statistic Finland's multidimensional database tables are distributed through a PX-Web interface. In early 2015, a totally new version of PX-Web was also taken into use. Some important elements in PX-Web have changed and because of that richer metadata, such as defining time variables and totals for the classification variables etc., is required for the individual database tables from now on.

1.2. Roles of the different release elements

The training also discussed the different release elements such as texts, graphs, tables and database tables, and highlighted their specific roles as conveyors of knowledge. It is important to understand that different release elements answer different needs. Statistical graphs are a way to disseminate quick mental images of phenomena behind numbers. Statistical tables, on the other hand, are used to represent those exact numbers behind graphs, in some cases the same phenomenon through many different units or in some cases more multidimensional structures than can be presented with graphs. In other words, they are for further use of those exact numbers. Database tables, in turn, enable data mining from larger data structures. They can be used for finding other interesting insights from the statistics than what was highlighted in the release or to create individually interesting presentation tables. In other words, they are used for all kinds of further processing of more extensive data.

It was also emphasised how the use of correct graph types and presentation methods has a significant effect on the understandability and usability of statistical data. It is very important to understand that each graph type also answers a different need. Different graph types are not straightforward substitutes for each other but reveal different sides and views to data. In most cases the data itself defines exactly what graph types and methods should be used. All the different graph types also have their own recommended presentation methods, which have to be taken into consideration. For example, when the graph axis can be cut and when it can't? When sorting has the desired effect and when not? When you have too many lines, bars, slices and so on.

The sole purpose of statistical graphs is to convey an image of the phenomenon behind numbers. So wrong graph types and presentation methods convey the wrong knowledge. So it can be said that a bad graph is far worse than no graph at all. As a statistical organisation, we have a responsibility to use graphs objectively and according to our best knowledge. That should be the standard that sets us apart from, for example, marketing firms and other less objective data producers. Moreover, individual graphs also have to follow standardised presentation forms (such as fonts, colour scales, line thickness etc.) in order to create a uniform style of graphs and for their part reinforce the unique visual design of Statistics Finland.

Also with tables, the right presentation methods are fundamental. For example, different positioning of the variables and values often radically changes the focus of the table. And a badly made table can easily blur all the relevant information and make the table extremely hard to read.

1.3. Role of the statistical experts

The aim was also to inspire experts to consider their role in conveying information: to summarise the most essential information from vast data masses. Their insight can make statistics significantly more understandable to our non-professional customers. Their expertise is needed to highlight the most interesting characteristics of the data, which in fact is the basic purpose of the whole release. Monthly or quarterly releases are typically slightly different from annual releases. In these, the customers often rely more on the internal stability and standardised presentation of the releases; same strategically important figures, graphs and tables presented in the same way as before.

1.4. Changing data environment

One question commonly raised at the training was, why do the instructions change so often? The answer is of course our constantly evolving data environment. The most important being the still continuing shift from paper publications to electronic originals. Recent economic development has created considerable pressure for public-sector organisations such as statistical offices. Due to decreasing resources, more and more automatic reprocessing methods are called for and statistical production processes have to be streamlined as efficiently as possible. It has become very important to utilise all kinds of API solutions and machine processing methods in our statistical production. Also, our customers will utilise API solutions more often next to the direct use of the PX-Web user interface. So from now on, data has to be in a machine-readable format which means highly standardised structures and metadata.

Standardised metadata is the bread and butter for automatic reprocessing of data. It creates a solid foundation on which all the API-based products and services can be safely and efficiently constructed. Without standardised metadata, significant amounts of manual work and tailoring are required and human errors cannot be avoided. The exception does not prove a rule for a machine. If the file names, variable names or value codes change from release to release all the automatic queries, links etc. break down. Or, if these names and codes do not make an exact match, the information from different tables cannot be united automatically. Most important of these variables are the ones that are commonly used in many different statistics such as geographical and time-related variables. Because machine processing relies so heavily on these standardised metadata, the training strongly emphasised their critical importance in the future.

One aim was also to make the PX-Web user interface more familiar to our own people, so if needed, they can also help our external customers to appreciate its full potential. It was highlighted that all data related to release should be put into StatFin (our PX-Web database for all free-of-charge aggregated data) and thus be available to all for further processing. Everybody can make their own saved queries and create their own tables for their own specific needs from these database tables. It was also highlighted that the design of these individual database tables is very important and cannot be overestimated. Tabulation, sorting, clear and distinct spelling of variable and value names, relevant metadata … are all fundamental aspects for the usability of PX-Web tables. Most of the criticism we have received related to the PX-Web user interface is actually not the fault of PX-Web itself, but is a result of poor design and insufficient content of the individual tables themselves.

Moreover, all these tables should be trilingual (Finnish, Swedish, English) and specifically in trilingual-files if possible. It was further highlighted that all these px-files have to be validated (with PX-Edit) before publishing. That is the only way to ensure that they really work in PX-Web. It was also emphasised that structural changes to the database tables have to be reported, so that necessary changes can be made for our own API queries and that in future we can pass that information also to our external API users.

1.5. Conclusions and results

Before the training sessions all the recent releases (including html-tables, graphs and database tables) of the statistics in question were reviewed and evaluated. It was much more concrete, interesting and useful for compilers as their own tables and graphs were used as examples. The main aim of the training was to provoke experts themselves to critically consider the contents of their own releases. It is useful to keep in mind that all statistics have their own specific features and global rules and recommendations cannot always be applied. In many cases, the right presentation methods are best based on special characteristics of the statistics in question and special demands of our customers, which are best known by these experts.

One of the discoveries of this evaluation was that there tended to be too many html tables in the releases. It was also found that many of those individual tables (and also graphs) tended to be too large and contain too much information, which makes it very difficult to perceive the most relevant information, which in fact should be their main purpose as static html-tables can hardly be used for anything else. So one of the most relevant feedback was to shift the focus of the releases from the static html-tables to the database tables because they can be used in several different ways.

At least in some cases the structural changes of these releases, tables and database tables requires a lot of work and will take some time to finish. So in the short run, by far the best result of these training sessions has been that compilers do ask and consult "technical support" more often than before. Collaboration in general has evolved significantly, the reasoning behind recommendations was clarified to all and many unnecessary misconceptions were removed.

2. Writing better statistical releases

2.1. Structure of the text

In order to make the texts of statistical releases more uniform, Statistics Finland has compiled instructions on, for example, the structure, length and headings of the text. However, statistical releases differ very much from one another. Some are very short releases describing monthly indicators, some are long releases describing annual statistics. In the training, examples were used to illustrate what the structure of a good statistical release text is like.

The heading is the main element of the text: it makes the reader interested or uninterested in reading.

A good heading of a statistical release:

- Summarises the main topic of the release
- Highlights the most important, interesting, topical issue: the biggest, smallest, new, continuous trend, etc.
- Does not try to say everything
- Includes an active verb

The first paragraph should summarise the main issue of the release. It complements the heading and presents the main findings on the topic discussed in the release. The entire content of the release is not summarised in the paragraph, only one or two main points are selected. The first paragraph should be short, at most five lines long.

The text of a good statistical release should be written like a news article written by a journalist. The structure of news also works in statistical releases as "the inverted pyramid". Things are presented in the order of importance: new, most important information first. In the following paragraphs, the topic and statistical data are examined in more detail. Possible background information and methodological explanations are placed at the end of the release.

The text paragraphs should be relatively short: one topic is one paragraph. The recommended length of the paragraphs is at most five lines but variations in the length of the paragraphs give the text rhythm. One should, however, avoid listing information in the text: the text should point out the meanings and connections of things.

In a long text, sub-headings make it easier to skim through and read the text. It is recommended to add sub-headings every two to three paragraphs.

2.2. Clarity and comprehensibility of the text

In the training, special attention was paid to the clarity and understandability of the text. The text of releases should be clear and neutral standard language. Simple and short sentences should be favoured, i.e. one idea is one sentence. Long and meandering sentences make reading difficult. Listing of figures presented in tables should also be avoided in the text.

Releases should use descriptive language and active forms of verbs rather than passive ones. If key statistical terms must be used in text, they should be explained when they occur in the text for the first time.

Avoid:

- ”Elevator statistics”: this went up, this went down

- Abbreviations
- Professional terms (define the key terminology)
- Listing figures shown in tables

The training also included language guidance: for example, the most common grammar mistakes or sentence structure related problems found in statistical releases were discussed.

2.3. Requirements of online writing

Because new statistical data are always first published on the web at Statistics Finland, the writers must also consider the requirements of writing for the web.

Many writing instructions, like those related to the news-like structure of the text and headings also apply when writing for the web. However, for example, the importance of the headings and the browsability of the text become emphasised when writing for the web.

Headings become particularly important in online texts as search engines rank them highly in search results. The headings of statistical releases are also visible in several lists on Statistics Finland's website and are automatically shown on Statistics Finland's Twitter account.

On the web, the length of the text affects readability. At Statistics Finland, the recommendation is that the length of a statistical release should be at most two screens (or two pages in PDF format).

2.4. Media as conveyors of statistical data

The role of the media as conveyors of statistical data was also discussed in the training, as the release texts are widely used in the rapid news flow of the web media.

On average, 200 news items a week are published in the web media on Statistics Finland's statistical releases, in the printed media, there are, on average, 100 news on statistical releases per week. Especially in the web media, the statistical release texts are often published fully or partially in their original format. Often reporters only change the headline for the news.

The clarity and understandability of the release text become emphasised in such rapid news flow: the risk of misunderstanding grows if the text is difficult to understand.

2.5. Continuous training and support

We have already previously offered compilers of statistics support in producing figures, tables and text. Such inquiries have clearly increased as a result of the training.

A total of 16 training sessions were arranged over a two-year period and 170 compilers of statistics participated.

Further measures related to the training are currently being planned. Thus far, language guidance training by an external expert has already been agreed. We are also planning "annual checks" of statistics, where we could together with the compilers of statistics go through the problems in the releases.