Country report for Finland

Statistics Finland   www.stat.fi

The Statistics Finland Web Portal Project:
All national statistical data and statistical services are available from a single source.
The data provided by Statistics Finland are widely used by the public and the private sector, linking and combining them with other data sources, creating new services and knowledge in the process.

Goals for the Project:
1) Remodel the structure of the website to be more user-driven for ease of use.
2) Update the content to fit changing user needs of the future
3) Modernise the user interface of the web site.
4) Reform the content production and publication procedure and systems to improve flexibility, speed and correctness of publications.

Statistics Finland's new concept for publishing statistical information through the main web portal relies on PxWeb and PxGraf on four important use cases:

1) PxWeb data cubes containing the aggregated statistical data will be listed, grouped, searched and presented to end users based on the various metadata of the data cubes and the relevant context in the web portals, ignoring the hierarchical folder structure and location of the actual px-files.
2) Key values, graphs and tables that automatically show the latest available values from PxWeb data cubes will be more ubiquitous part of the new web portals. Graphs and tables will also support limited browsing of different dimensions and time periods, lessening the need to publish many very similar tables and graphs.
3) Automatically updated graphs, tables and key values will be supplemented by "frozen" snapshot versions of them that will be used in and archived with statistical publications.
4) The aim is to have most (if not all) published data available through PxWeb and to make updates to PxWeb data cubes as visible to the end users as the actual statistical publications.

The PxPro project:
Statistics Finland has developed the PxPro system for producing (continuously updating) px-files. PxPro consists of a metadata editor, used to create and maintain tables' metadata, a corresponding metadata repository, and a new SAS-macro that produces a px-file by combining the data from a

International PX Reference Group meeting 2019

SAS file with metadata from the repository. This system opens several possibilities for automation and standardisation:
- if possible, values for metadata elements are restricted to lists
- if possible, px-keywords are generated automatically
- errors are discovered and reported to the user as early as possible, etc.

To produce a px-file with PxPro, the user first needs to define the table metadata such as the table title and subject area as well as each variable's type, label, description etc. Each grouping variable needs to be ascribed a reference to a classification. The classification contains the information on the correspondence of each group's code to its multilingual labels. The classification database is external to the PxPro system and serves other purposes as well. All the metadata defined with the editor are stable and do not need to be modified with each publication. To produce a px-file, the PxPro SAS macro is run with a small number of parameters.

Problems and/or good news
PxWeb and the px-family programs have been performing beyond our expectations. PxWeb 2019 v1 has been especially robust.

Which version of PX-Web do you use?
All our 12 PxWeb servers use the latest PxWeb 2019v1

Do you use the API?
The PxAPI is in use in all our PxWeb servers, including the chargeable PxWeb services (with PxGatekeeper). We have been teaching the API usage as part of our database courses.

Any reactions from users of interest?
The feedback has been mainly positive.
PxAPI2.0, maps, good graphs and a mobile interface will be greatly welcomed by the users.

Update to new version?
We always update our servers to the latest version of PxWeb, if it passes our tests.

Development

PxGraf
Statistics Finland has continued to develop PxGraf which will be used on our new future website.

International PX Reference Group meeting 2019

PxEdit and PxJob
The new versions (4.0) of PxEdit and PxJob will be available at the meeting. Its main new feature is the extended structural text format among other smaller new features, changes and speedups.

PxGatekeeper (PxWeb user rights control)
The latest version now also supports saved queries and the PxAPI (IP address or username/password).

Something you would like to show or discuss at the meeting?

Show: new version of PxEdit, working demo of PxGraf.
Discuss: possible changes to the px file format.

Contact persons

Hans Baumgartner	hans.baumgartner@stat.fi	
Veli-Matti Jantunen	veli-matti.jantunen@stat.fi	[PxEdit & PxJob]
Kim Huuhko	kim.huuko@stat.fi	[PxGraf]
Rok Platinovsek	rok.platinovsek@stat.fi	[PxPro]

Appendix 1

Organisation-specific metadata extensions
Working with px-files, we have sometimes noticed the need to define metadata that are beyond the scope of the px-format. One example are metadata that could be used to enrich PxGraph visualisations: a measure's theoretical minimum, maximum and baseline value (e.g. an index's base year) could be added as reference lines to a plot. Adding new keywords to the px-format, however, usually takes time as it demands the px-community to agree to the suggested syntax and make appropriate modifications to software and databases. A quick solution one is tempted to make in such cases is to (mis)use an existing px-keyword that is not used by their organisation to accommodate the new metadata.

In the case of PxGraph's metadata needs, Statistics Finland is considering (mis)using the META-ID keyword as follows.
```
META-ID("information", "percentage")="min-value=0, max-value=100, baseline=";
```

To avoid such quick-and-dirty solutions, we propose extending the px-format with keywords that allow an organisation to add its own metadata to px-files.

One possible way to declare measure-level attributes would be as follows.
```
MEASURE-ATTRIBUTES="min-value","max-value","baseline";
MEASURE-ATTRIBUTE("min-value","percentage")= "0";
MEASURE-ATTRIBUTE("max-value","percentage")= "100";
```

In the above example, the keyword MEASURE-ATTRIBUTES declares the measure-level attributes used in the px-file. The MEASURE-ATTRIBUTE keyword's first sub key then refers to the attribute and the second sub key to the measure (a measure is a value on the dimension marked with CONTVARIABLE).

Sometimes the metadata we would like to add to px-files do not concern measures as in the example above, but rather dimensions or the entire table. Equivalent keywords (DIMENSION-ATTRIBUTES, DIMENSION-ATTRIBUTE, CUBE-ATTRIBUTES, CUBE-ATTRIBUTE) could be used to declare dimension-level and table-level attributes.

These additional keywords would allow organisations to include in the px-files metadata beyond the scope of the px-format without resorting to misusing exiting px-keywords. Such organisation-specific metadata could be used, e.g., when developing new features such as the aforementioned PxGraph. If the new metadata are found to be beneficial for the entire px-community and once their use has matured, they could be "canonised" as px-keywords. Otherwise they would retain their organisation-specific status.

JSON-stat's extension object seems to address the same need: *extension allows JSON-stat to be extended for particular needs. Providers are free to define where they include this property […].*

Multiple content dimensions

In comparison to organisation-specific extensions, the issue of tables containing multiple content dimensions is of secondary importance to Statistics Finland. We will briefly discuss multiple content dimensions, however, because they should be considered in the syntax for the organisation-specific extensions.

The px-file format currently allows only one dimension to be marked as the content dimension (CONTVARIABLE). The assumption of a single content dimension seems to be too rigid, however, considering how PX-tables are used in practice. Several Statistics Finland's tables can be said to contain multiple content dimensions, the most prominent example being national accounts' publication tables. Similar tables can be found in PX-databases of other Nordic statistical offices; see links: Finland, Denmark, Norway, Sweden.

This might seem as a slight misuse of the model, but it has implications for machine readability. Consider, e.g., that we want to develop a visualisation application that asks the user to select a measure from a pull-down list, then produces a plot for the selected measure. If the cube in question contains two content dimensions, then two pull-down menus must be generated to select e.g. "Exports of goods" and "at year 2010 prices".

Indeed, the cube could be "flattened" by constructing the combinations of all content dimensions' values and representing the combinations as values of a single content dimension. This would, however, obfuscate the inherent structure of the table (that measures can be thought of as combinations of dimensions and that all such combinations are possible), and likely result in a high number of combinations making it difficult for the user to find the one of interest.

In order to accommodate multiple dimensions, the value of the CONVARIABLE keyword could be allowed to be a list. Alternatively, a new keyword could be introduced to avoid confusion.

```
MEASUREVARIABLES("Transaction","Information");
UNITS("*","Current prices")="milj. eur";
UNITS("*","Changes in values, %")="%";
LAST-UPDATED("*","Current prices")="20190315 09:00";
LAST-UPDATED("P71R Imports of goods","Current
prices")="20190315 09:00";
LAST-UPDATED("P71R Imports of goods","Changes in values,
%")="20190415 09:00";
```

In the above example we illustrate how the asterisk notation could be used to declare default values for certain measure attributes.

We note that JSON-stat already allows for multiple content dimensions via the metric array. *It can be used to assign a metric role to one or more*

dimensions. It takes the form of an array of dimension IDs in which order does not have a special meaning.

Implication for organisation-specific metadata extensions
The syntax for organisation-specific metadata extensions should be extendable to multiple content dimensions. In the below example the first sub key of MEASURE-ATTRIBUTE refers to the attribute and the remaining sub keys identify the measure via values of the content dimensions (dimensions marked with CONVARIABLE/MEASUREVARIABLES).
```
MEASURE-ATTRIBUTES="min-value","max-value","baseline";
MEASURE-ATTRIBUTE("baseline","*","Changes in values, %")= "0";
```

Identifiers
Statistics Finland has developed and will shortly be adopting a GSIM-based metadata system. One of the system's aims is to trace variables from data collection to dissemination. Variables will also be categorised with regard to content. We intend to use this metadata on Statistics Finland's new website (under construction) to improve the discoverability of px-tables. We therefore intend to include in the px-files identifiers for variables (measures and dimensions), classifications, variable categories and even certain methods used in computing the measures (indexation, deseasonalisation). If possible, we would like to include this information in px-files without resorting to misusing exiting px-keywords.

For details see the NSM conference paper on the subject.

continues …

Appendix 2

Wish list:

- TIMEVAL-definition should work without any unnecessary "parameters".

- Saved query (and API) must always use "ALL" definition in cases where every value is to be selected. "Empty" definition must be reserved to ELIMINATION purpose only.

  - So in cases user has selected all individual values from a variable, saved query generator has to ask the user does he mean "ALL" or just those values he just selected … and the saved query must be generated accordingly.

- Variable search (selection window) function would be much more useful, if it would include basic logic operators as AND and OR etc. And some kind of tooltip for that would be very useful!

- Variable code to a px file format. Now is the time!

- New keyword for defining order of tables only.

  - MATRIX, file name and TITLE has their own separate purposes which are not to be messed with order of tables.