# Toward standardized statistical tables: metadata-driven px-file production

Rok Platinovsek, Statistics Finland, [rok.platinovsek@stat.fi](http://intranet.stat.fi/teams/2241-px-pro-jalkautus-ja-jatkokehitys/Jaetut%20asiakirjat/NSM2019/rok.platinovsek%40stat.fi)

Daniel Davis, Statistics Finland, daniel.davis@stat.fi

Emilia Saarinen, Statistics Finland, [emilia.saarinen@stat.fi](http://intranet.stat.fi/teams/2241-px-pro-jalkautus-ja-jatkokehitys/Jaetut%20asiakirjat/NSM2019/emilia.saarinen%40stat.fi)

**Abstract**

Statistical tables (aggregated datasets) are arguably the most important content published by national statistical institutes (NSIs). When the production of statistical tables in the NSI is decentralized, however, their content may vary across statistical programs (the same variable may, e.g., appear with different names in different tables).

Like all other Nordic NSIs, Statistics Finland uses the PX-suite for the production and dissemination of statistical tables. Statistics Finland is preparing a major renewal of its website that will base content like key statistical figures and visualizations directly on PX-tables. To fulfil the new website’s requirements, the content of our statistical tables needs to be standardized and the quality of their metadata needs to be improved significantly. To facilitate this transition, Statistics Finland has developed PX-Pro, a system comprised of two tools. An editor is used to maintain a statistical table’s metadata while a SAS-macro is used to produce the corresponding px-file.

We discuss the PX-Pro system with its relations to existing information models like GSIM, the Data Cube Vocabulary, and the Nordic Data Model (CNMM) as well as our experience with PX-Pro’s introduction to statistical production. So far, users have received the new tools very positively. The challenge, instead, has been with developing standards that have thus far been lacking, e.g. harmonizing classifications across statistical programs. We propose a way of standardizing the names of variables by referring to certain GSIM objects the variables are related to. Another challenge we discuss are tables with multiple content dimensions that according to our understanding do not fully comply with the px-file format yet are widely used by NSIs.

The paper highlights concrete benefits of using GSIM in the standardization of statistical tables. The solutions we propose can be readily utilized by other NSIs.

**Keywords**: Open data, PX-Web, standardization, GSIM, metadata

## Introduction

Statistics Finland is currently undergoing a modernization of many of its production systems. The renewal most visible to our customers will be a new website that will base content like key statistical figures and visualizations directly on statistical tables as well as significantly improve discoverability by interlinking related content. This imposes demands of a new kind on the statistical tables. They must be made machine-readable and standardized across statistical programs.

As it might not be immediately apparent what machine-readability means in this context, let us illustrate with an example. Say we have the following description of a statistical table that we want to make machine-readable.

The table *Index of turnover in industry* reports values on the *Turnover index* and the *Trend of the turnover index* by *Year* and *Industry (NACE)*.

A human can intuitively understand what objects the given information pertains to, especially when presented with the concrete table in question, but a computer cannot. To render the information machine-readable, we must make explicit that *Turnover index* and *Trend of the turnover index* are labels of measures and *Year* and *Industry (NACE)* labels of dimensions. The table’s measures and dimensions might have other properties that must also be given explicitly. The information, e.g., that the *Year* dimension pertains to time can be grasped from its label by humans but must be made explicit for computers, e.g., by stating that for the dimension in question *role=”time”*.

As mentioned, we want to interlink content to improve discoverability on the new website. This means, e.g., suggesting to a visitor examining a particular statistical table, that they might be interested in other tables containing the same measures or dimensions. To allow this, the measures and dimensions must be marked with the same identifier in all tables they appear in.

As the way statistical tables are currently produced at Statistics Finland relies on ad-hoc manual input of metadata for each table, it became immediately apparent that the aforementioned goals could not be achieved without a modernization of the process. To address the issues with the production of statistical tables, Statistics Finland has developed the PX-Pro system that is the topic of this paper.

We begin by outlining the terminology and basic concepts behind statistical tables in Section 2, proceed to describe the PX-Pro system in Section 3, and conclude by outlining a number of future development possibilities in Section 4.

## Terminology and concepts

The Data Cube vocabulary (W3C, 2014) is a useful conceptualization of what constitutes a statistical table. According to the Data Cube vocabulary, a statistical table (or any dataset for that matter) is a collection of observations that can be characterized by a set of dimensions that define what the observation applies to (e.g. time, geographical region, gender) along with metadata describing what has been measured, how it was measured, and how the observations are expressed (e.g. measurement unit). The cube is organized according to a set of dimensions and measures. Dimensions components identify the observations: giving a value on each dimension identifies a single observation. The measure components, in turn, represent the phenomenon being observed. Multiple measures can be represented within the same cube by introducing an additional dimension whose value indicates the measure being conveyed by each observation.

Another framework we want to draw upon is the Generic Statistical Information Model (henceforth GSIM; see UNECE 2019). GSIM is an internationally endorsed framework that provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics. As such, GSIM plays an important part in modernizing and aligning the standards associated with official statistics at an international level. We will only mention GSIM’s objects pertinent to our discussion of statistical tables.

GSIM recognizes several abstraction levels of what may informally be termed “variable” as well as corresponding entities defining classes of objects (see Figure 1). At the most concrete level, the Population is used to describe the total membership of a group usually using temporal and geographic boundaries, e.g., *University students residing in Finland on 31.12.2018*. One level of abstraction higher is the Universe, defined as a class of people, entities, events, or objects, with no specification of time and geography (*university students*). More abstract still is the Unit Type, which is simply the class of objects of interest (*person* in our running example).

Figure 1: GSIM entities pertinent to statistical tables.



GSIM recognizes three abstraction levels of variable corresponding to the Unit Type, Universe, and Population. At the most abstract level, the Variable is defined as measuring a certain characteristic with no indication of how the characteristic in question is to be represented in the data set. Specifying how the Variable will be represented yields the Represented Variable. This requires us to establish whether the variable will hold numeric values and, if so, what range of values is allowed and what measurement unit is to be used. On the other hand, the variable might be represented by character codes, in which case the set of allowed codes as well as their meanings is specified. These characteristics are ascribed to the Represented Variable via the Value Domain. At the most concrete level, the Instance Variable constitutes the use of a Represented Variable within a data set. It is the Instance Variable that holds concrete values in its data points.

Figure 2: Annual turnover index table to illustrate Data Cube and GSIM entities.



GSIM is not explicit about how longitudinal aggregated datasets – statistical tables – are to be conceptualized, but our interpretation is as follows and is illustrated in
Figure 2. The table has three dimensions: *Year*, *Industry*, and the content dimension holding four measures. Since an Instance Variable measures a single population – pertains to a single time point – the shaded slice of the measure *Turnover index* corresponds to an Instance Variable. The entire column, in turn, corresponds to the Represented Variable. The whole time series does indeed, typically have the same Value Domain properties that are characteristic of the Represented Variable. The GSIM Variable object, on the other hand, is an abstraction beyond the statistical table’s structure. It often happens, though, that the table’s various measures measure the same property in different ways, in which case they can be said to be specifications of the same Variable (shown as boxes above the table in Figure 2).

## Statistical table dissemination and the PX-suite

Statistics Finland, like many other European NSIs, uses the PX-suite of programs for the production and dissemination of its statistical tables. Statistics Finland’s dissemination is *file-based*, i.e., files adhering to the px-file format (Likidis 2013) are generated and placed into a database. A program called PX-Web can then read a file’s content and generate a webpage with the statistical table’s content that has been requested by a browser. Generating a px-file can also be circumvented if PX-Web is used with an SQL-database. In that case both a table’s data and metadata are saved in an SQL-database structured according to the Nordic Data Model (CNMM; see Statistics Sweden 2019). PX-Web can then be configured to generate the webpage content directly from the SQL-database without an intermediate px-file phase. Statistical offices using the CNMM have typically developed their own applications for maintaining the tables’ metadata and loading data into the SQL-database.

### The motivation behind PX-Pro

Statistics Finland’s statistical table dissemination is decentralized: px-files are produced by the statistical program who is responsible for their content. While px-files can be produced with several software tools, the most common way involves running a SAS macro that takes an input SAS dataset and produces the corresponding px-file. The metadata required to document the statistical table are hard-coded into the SAS-code used to produce the px-file. This way of producing statistical tables is problematic for the following reasons.

* Maintaining metadata hard-coded into the SAS code is error-prone. Possible errors cannot be reported to the user as values of metadata filed are changed but are detected further down the pipeline.
* No distinction is made between stable metadata and publication-specific metadata like the number of preliminary data periods and the date of last update.
* It is difficult to enforce standards across statistical programs, as users can freely type their own values into the metadata “fields” as opposed to choosing them from a predefined list.

Attempting to establish standards, the department coordinating the dissemination of statistical tables must rely on written instructions that can be misunderstood or disregarded. If a new standard is to be enforced or an old one changed, all users must make according changes to their SAS code. This process is so inefficient that only the most important pieces of metadata in Statistics Finland’s publication tables have been attempted to be standardized.

### Architecture and usage

To address the issues with the production of statistical tables, Statistics Finland has developed the PX-Pro system. PX-Pro 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 SAS file with metadata from the repository (see Figure 3). This opens several possibilities for automation and standardization: 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, the user first needs to define the table’s metadata such as the table’s 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 PX-Pro system and serves other purposes as well.

Figure 3: PX-Pro’s architecture.



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:

* the name of the SAS dataset containing the table data,
* the corresponding table ID used to retrieve the table’s stable metadata,
* parameters giving publication-specific metadata like the number of preliminary data periods and the date of last update for the measures.

The macro retrieves the table’s metadata from the repository and fetches classifications from the classification repository. Variable names are used to link each variable to its metadata and grouping variables’ codes are used to link them to their labels in the classification. The macro run is aborted if any discrepancies between metadata and data are encountered, e.g., if a code is found in the data that does not have a counterpart in the classification. If no discrepancies are encountered, a px-file is produced. The px-file then needs to be manually transferred to the database wherefrom PX-Web can access it.

The PxPro SAS macro automates several tasks that previously needed to be done manually, including the following.

* The start and end point of the time series are appended to the table’s root title.
* Links to the statistical program’s web-pages (such as the methodological documentation, concepts and definitions etc.) are appended to the footnote for existing web-pages.
* Footnotes are generated explaining that time periods marked with an asterisk are preliminary.
* A legend is appended to the footnote explaining the meaning of special symbols (various types of missing data) in the table’s data.

In addition to reducing the user’s burden and ensuring that the content is correct, the automation results in content standardized across statistical programs.

We note that certain tables contain a single measure or pertain to a single time point in which case the author of the table might want to mention the time point or the measure’s content only in the table’s title making the information non-machine-readable (since the computer does not know to look for it there). To avoid this, we have made the use of both the time dimension and the content dimension mandatory in PX-Pro. In case of a single time point or single measure, the corresponding dimension will contain a single value.

### Adoption of the PX-Pro-system

In comparison to the old SAS-macro, the PX-Pro system greatly simplifies the production of statistical tables. Early adopters have been quite enthusiastic about making the transition to the new system. At the time of writing, 77 statistical programs have transitioned to using PX-Pro constituting half of Statistics Finland’s primary PX-database StatFin. The remainder of the statistical programs will need to transition to PX-Pro to comply with the new website’s requirements as it comes online.

The adoption has been complicated by the fact that statistical programs differ widely with regard to how they produce statistical tables. Some tables were previously produced with little or no automation using a spreadsheet editor and PX-Edit (Statistics Finland, 2019). Migrating the production of such tables to the PX-Pro system has been much more challenging than tables that were already produced with SAS.

The greatest encountered challenge, however, does not relate to software at all, but instead concerns statistical classifications used in the dissemination tables. Previously, each statistical program was responsible for creating and maintaining its own classifications, typically resulting in classifications that were not suitable for other statistical programs. The project introducing the PX-Pro system has therefore expended a great amount of effort to foster discussions across statistical programs regarding their shared content resulting in the creation of harmonized classifications[[1]](#footnote-1) that can be used across statistical programs.

## Future development

We have thus far introduced the PX-Pro system and outlined how it is used to standardize the content of statistical tables. We still see much room for improvement, however. The following sections give our thoughts on how tables’ content can be further standardized and enriched with additional metadata.

### Tables with multiple content dimensions

The px-file format specification (Likidis, 2013) documents metadata that can be defined at table-level, variable-level, or cell-level. Properties like the measurement unit and date of last update are, by default, table-level properties under the assumption that all the table’s measures share them. Since the px-file format was first specified, this was found to be insufficient and the format was expanded. Marking the content dimension with a special px-keyword endows it with a special role in the cube: its values identify the measures. This allows measures to be given different values on the measurement unit, date of last update and several other properties.

The px-file format allows only one dimension to be marked as the content dimension. The assumption that there is only one content dimension in the cube seems to be too rigid, however, considering how PX-tables are used in practice. Several Statistics Finland’s statistical tables can be said to contain multiple content dimensions, the most prominent example being national accounts’ publication tables. One such example[[2]](#footnote-2) is given Figure 4.

Figure 4: Screenshot showing variable selection in PX-Web for the table Gross *domestic product and Gross national income*.



A measure in the example table does not correspond to the value on one dimension but arises as the combination of values on two dimensions (*Transaction* and *Information*), e.g., *Exports of goods* and *at year 2010 prices*. The value of the former dimension denotes the measure’s content while the value of the latter dimension gives the method by which the measure was calculated.

This is indeed a practical way to organize this table. If we were to restructure the table to make it fully comply with the px-file format, we would need to “flatten” it by constructing the combinations of all content dimensions’ values and representing the combinations as values of a single content dimension. This would have several negative consequences.

* The number of constructed combinations would be prohibitively large in certain tables.
* The inherent structure of the table – that measures can be thought of as combinations of elements and that all such combinations are possible – would be lost or obfuscated.
* Any hierarchical relationships between values of the content variable would be lost or obfuscated. An example of such hierarchical relationships in the National Accounts table is that the category *Imports of goods and services* is comprised of *Imports of goods* and *Imports of services*.

Instead of flattening existing tables to comply with the px-file format we propose the format be expanded to explicitly allow several content dimensions. We give an example of how the format could be expanded in Appendix A. Explicitly marking all content dimensions is critical for machine-readability. When a graphics engine, e.g., generates a time-series plot of a table’s measures, it must be able to identify all content dimensions to label the measures correctly.

### Standardizing measures’ labels

As the production of statistical tables at Stats Finland is decentralized, their content has been found to vary across statistical programs. The same seasonally unadjusted measure can, e.g., be referred to as the *Original index series* in one table and the *Point figure* in another; the index’s base period may appear in the measures’ names in one table and in the table’s title in another etc. One goal for the future is therefore to extend standardization to the names of measures.

In an informal survey of Statistics Finland’s tables, we tried to determine whether statistical programs employ any patterns when naming measures. We have found that measures’ names seem to consist of a limited number of elements.

1. The measure’s content.
2. The name of the summary statistic that was used to aggregate the microdata to the table’s level, e.g., sum, mean, 1st quartile, median, 3rd quartile.
3. Statistical methods applied after aggregation to adjust the variable in question or facilitate comparison to a different time point. These include the following categories.
	1. Seasonal adjustment methods, e.g. the terms *seasonally adjusted*, *working day adjusted* or *trend* may appear in the variable’s name.
	2. Indexation, e.g. the Laspeyres or Paasche index relative to a base period.
	3. Change percentages, e.g. annual change, monthly change or the three-month moving average percent change.
4. The measurement unit.

We have not encountered a measure whose name consisted of all these elements. Rather, an element appears in the measure’s name when it is required to differentiate between measures in the context of the table. When, e.g., several summary statistics (median, 1st quartile, 3rd quartile) are reported for wages, the aggregation method appears in the measures’ names, otherwise it is usually omitted.

The above categorization provides a basis for the standardization of the measures’ labels. This would involve modifying the PX-Pro editor and SAS macro so that instead of typing a measure’s label, the user would provide the information on the variable’s content, measurement unit and methods involved in the variable’s calculation: a known number of elements each with a controlled vocabulary. The variable’s label could then be formed by concatenating the elements’ labels. This would, of course, yield a more “mechanical” sounding label, e.g., the standardized form of the label *Annual change of the new orders in manufacturing index* would be *New orders in manufacturing, index (2015=100), annual change*.

The most difficult element to standardize pertains, of course, to the measure’s content. According to our understanding, the GSIM Variable could serve well as the element capturing the measure’s content. Other projects underway at Statistics Finland are developing a GSIM-based metadata system that will allow GSIM Variables to be documented (see Kaukonen and Saloila 2019).

While our main motivation has been to standardize measures’ labels, providing information on the content, methods of computation, and measurement unit for each measure has other benefits. This information can be included in the px-files in the form of identifiers, interlinking tables with similar content (see Appendix B). It also makes it possible to automatically generate a footnote for each measure with links to web-pages detailing the measure’s content as well as descriptions of the methods used in the measure’s computation. These descriptions can be maintained by substantive or methodology experts, making their content substantially better in quality and more exhaustive than if it written by the authors of tables themselves (the present state at Statistics Finland).

Finally, as GSIM Variables are introduced, linking them also to tables’ dimensions would be very beneficial. At present, the customer wanting to find all tables containing, e.g., the *economic activity* dimension would need to know all labels it was given in different tables. Linking instances of economic activity to the same GSIM Variable would enable its label to be standardized across tables and allow the tables sharing this dimension to be discovered in searches.

### A note on measurement units

While the px-file format specifies how to technically declare the measurement unit, it gives no indication of its content. Examining Statistics Finland’s publication tables, it became apparent that this piece of metadata is not synonymous with how the measurement unit is defined in physics. A count, e.g., is a unitless quantity while proportions, percentages, and indices are calculated in a way that cancels out their measurement unit to unity. Still, in PX-tables these quantities are routinely ascribed measurement units, e.g., *number of persons*, *%*, *per mille*, and *index point*.

To standardize measurement units’ labels, we have therefore proposed to use the physical unit of measurement when such a unit exists. In the case of unitless quantities, on the other hand, the measurement unit should give information of the variable’s scale. The value should, e.g., sound reasonable when used to label the axis of a diagram showing the measure in question. When the quantity in question is a count, we have decided to include in its label GSIM’s Unit Type to provide additional information on what the count pertains to (e.g. *number of persons*, *number of occurrences*).

### Establishing comparability

At present, the px-file format specifies no information on the population whose characteristics are captured by the table’s measures. The absence of structured information on the table’s population hinders comparisons across tables as the person wishing to make such comparisons must examine the tables’ methodological documentation to establish whether the measures of interest are comparable. While the methodological documentation should, indeed, be consulted to unequivocally determine comparability, we suggest it would be beneficial to add population-related metadata to the tables. Including such metadata would also interlink tables referring to similar populations.

The concept of population is central to establishing comparability as relationships between a pair of measures can only be examined without additional assumptions when the measures in question pertain to properties of the same population. Because the population is specific to a time point, however, defining it at the measure level is not possible for a longitudinal data set like the publication table as we have pointed out in Section 2.

Following GSIM, the Population emerges as the Universe is combined with information on geography and time. As publication tables typically contain a time dimension, what is still missing to establish comparability is the information on the universe and geography. These could be added to px-files in the form of measure-specific keywords, as suggested in Appendix B. If values of both universe and geography were restricted to a controlled vocabulary, the customer could check that measure A from a particular table has the same values on universe, time, and geography as measure B from another table and thereupon establish that measures A and B are likely to be comparable.

## Conclusion

Statistics Finland has begun a shift toward more standardized, machine-readable, and interlinked statistical tables. The introduction of new software plays an important role in this transition as the new tools allow limiting metadata values to controlled vocabularies, perform consistency checks, and automatically generate content when possible. This paper has described the current state of statistical table production at Statistics Finland and highlighted some future ways of standardizing and enriching tables’ metadata we are considering. The most important of these is the interlinking of tables with references to underlying GSIM Variables that promises to significantly improve the discoverability of content on the new website.

The general theme of standardization we have pursued in the paper extends beyond statistical tables, however. As a visitor of the statistical office’s website moves from one page to the other, they should see content presented in a uniform fashion so that they can effortlessly find the information they are looking for. Such a consolidation of content will not be possible if we are not able to move beyond compartmentalized production in which each statistical program handles its own affairs in isolation from others. Software tools are important in this process as they allow the authors to pay less attention to the technical implementation and concentrate on the content instead. But the aim of content consolidation cannot be achieved merely by introducing new software without a shift toward shared content harmonized across statistical programs.

**Acknowledgments**

The authors would like to thank Veli-Matti Jantunen, Hans Baumgartner, Kim Huuhko, Kari Djerf, and Ragnhild Rein Bore for their comments on the draft version of the paper.

## References

Kaukonen, E. and M. Saloila (2019) Modernization of process, system and quality management, Nordic Statistical Meeting, Helsinki, 26.-28.8.2019.

Likidis, P. (2013) PX-file format specification, AXIS-VERSION 2013. Available at: <https://www.scb.se/Upload/PC-Axis/Support/Documents/PX-file_format_specification_2013.pdf> (Accessed: 15 May 2019).

Statistics Finland (2019) PX-Edit. Available at: [https://www.stat.fi/
tup/tilastotietokannat/px-tuoteperhe\_en.html](https://www.stat.fi/tup/tilastotietokannat/px-tuoteperhe_en.html) (Accessed: 15 May 2019).

Statistics Sweden (2019) PxWeb with an SQL-database. Available at: <https://www.scb.se/en/services/statistical-programs-for-px-files/px-web/px-web-med-sql-databas/> (Accessed: 19 June 2019).

UNECE (2019) Generic Statistical Information Model (GSIM) v 1.2. Available at: [https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model](https://statswiki.unece.org/display/gsim/Generic%2BStatistical%2BInformation%2BModel) (Accessed: 15 May 2019).

W3C (2014) The RDF Data Cube Vocabulary. Available at: <https://www.w3.org/TR/vocab-data-cube/> (Accessed: 16 May 2019).

## Appendix A

This appendix gives three examples of abbreviated px-files to illustrate the points made in Section 4.1. The content given in the examples does not constitute a complete px-file as we have omitted certain mandatory PX-keywords for brevity[[3]](#footnote-3).

The first example corresponds to a small retrieval from the table represented in
Figure 4. Keywords STUB and HEADING give the names of the cube’s dimensions (as well as their default orientation) while the VALUES keyword lists each dimension’s levels. The measurement unit (UNITS) and time of last update (LAST-UPDATED) are defined at table-level. The measurement unit is, in this example, misused resulting in non-machine-readability: all measures are declared to be measured in millions of euros. That there is an exception to this rule can be deduced from the label *Changes in values, %*. While this kind of unstructured information can be noticed by an observant human, it is not machine-readable.

Example 1: px-file without CONTVARIABLE keyword.

STUB="Transaction";

HEADING="Year","Information";

VALUES("Transaction")="P71R Imports of goods","P72R Imports of services";

VALUES("Year")="2017","2018";

VALUES("Information")="Current prices","Changes in values, %";

UNITS="milj. eur";

LAST-UPDATED="20190315 09:00";

Declaring the *Information* dimension as the content variable (CONTVARIABLE) allows a different measurement unit and timestamp of last update to be declared for each of its levels as in Example 2.

.

Example 2: px-file with CONTVARIABLE keyword; keywords STUB, HEADING and VALUES as in Example 1.

CONTVARIABLE("Information");

UNITS("Current prices")="milj. eur";

UNITS("Changes in values, %")="%";

LAST-UPDATED("Current prices")="20190315 09:00";

LAST-UPDATED("Changes in values, %")="20190415 09:00";

As discussed, the measures in this table do not actually correspond to values on one dimension but rather arise as the combination of values on two dimensions: *Transaction* and *Information*. As the px-file format does not currently support multiple content dimensions, we propose it be extended as in Example 3.

Example 3: px-file with MEASUREVARIABLES keyword; keywords STUB, HEADING and VALUES as in Example 1.

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";

If the value of an indexed characteristic is shared by all levels of a dimension, the asterisk character could be used to shorten its declaration, as we have done for UNITS in Example 3. We have also used this shorthand to declare the time of update of measures with *Information* = *Current prices*, while using the flexibility of the extended notation to declare a different timestamp for the measure arising as the combination of *P71R Imports of goods* and *Changes in values, %*.

## Appendix B

This appendix gives examples of how characteristics of measures, dimensions, and cubes could be represented using the px-file format using a limited number of additional PX-keywords. Keywords supported by the px-file format should be agreed upon and serve the needs of the many organizations that use the PX-suite. The requirement that PX-keywords be generic makes it difficult for an organization to include its own specific metadata in px-files. The notation we suggest in this appendix would allow organizations to specify their own set of measure-specific attributes using generic keywords.

An important requirement is that the notation should support tables with multiple content dimensions (see Appendix A). Example 4 gives measure-specific attributes for a measure appearing in Figure 2. The below example assumes the cube contains two content dimensions.

Example 4: Keywords specifying measure attributes.

MEASURE-ATTRIBUTES="Variable","AggregationMethod","StatisticalMethod";

MEASURE-ATTRIBUTE-ID("Variable","Turnover index","\*")= "urn:stat.fi:variable:a0001";

MEASURE-ATTRIBUTE("Variable","Turnover index","\*")="Turnover";

MEASURE-ATTRIBUTE-ID("AggregationMethod","Turnover index","\*")= "urn:stat.fi:aggregationmethod:c0001";

MEASURE-ATTRIBUTE("AggregationMethod","Turnover index","\*")="Sum";

MEASURE-ATTRIBUTE-ID("StatisticalMethod","Turnover index","trend")="urn:stat.fi:statisticalmethod:d0001", "urn:stat.fi:statisticalmethod:d0002";

MEASURE-ATTRIBUTE("StatisticalMethod","Turnover index","trend")=
"Laspeyres Index", "Trend";

The first index of MEASURE-ATTRIBUTE-ID and MEASURE-ATTRIBUTE gives the name of the attribute in question. The remaining indices identify the measure via the content dimensions’ values. The values of MEASURE-ATTRIBUTE-ID are identifiers of the measure’s GSIM Variable, aggregation method, and statistical methods. The values of MEASURE-ATTRIBUTE keywords give corresponding labels fetched from metadata repositories. The keyword MEASURE-ATTRIBUTES lists allowed names of measure-specific attributes to allow the px-file to be validated.

Statistical tables are produced at Statistics Finland by using classifications that are maintained in a separate classification repository. As mentioned, we should also soon have available a metadata repository with information on GSIM Variables. Example 5 illustrates how we would like to include this information in px-files to interlink tables with similar content.

Example 5: Keywords specifying dimension attributes.

DIMENSION-ATTRIBUTES="Variable","Classification";

DIMENSION-ATTRIBUTE-ID("Variable","Industry (NACE)")= "urn:stat.fi:variable:a0002";

DIMENSION-ATTRIBUTE("Variable","Industry (NACE)")="Industry";

DIMENSION-ATTRIBUTE-ID("Classification","Industry (NACE)")=
"urn:stat-fi:meta:classification:toimiala\_83\_20000101";

DIMENSION-ATTRIBUTE("Classification","Industry (NACE)")=
"Standard Industrial Classification (TOL 2008)";

Similarly, Statistics Finland’s specific needs to tag tables with information on topic and subtopic are addressed in Example 6.

Example 6: Keywords specifying cube attributes.

CUBE-ATTRIBUTES="Topic","Subtopic";

CUBE-ATTRIBUTE-ID("Topic")="urn:stat.fi:topic:f0001";

CUBE-ATTRIBUTE("Topic")="Industry";

CUBE-ATTRIBUTE-ID("Subtopic")= "urn:stat.fi:subtopic:g0001","urn:stat.fi:subtopic:g0002";

CUBE-ATTRIBUTE("Subtopic")="Manufacturing","Turnover";

1. The NACE classification code *Total industry*, e.g., referred to categories B through D in some tables and B through E in others. The agreed upon standard is now for *Total industry* to refer to B through E and to always prefix the label with the code BTE (meaning B-to-E). [↑](#footnote-ref-1)
2. A web search reveals that National Accounts tables of other Nordic statistical offices’ also feature multiple content dimensions and are organized in a similar manner. [↑](#footnote-ref-2)
3. Note that we are also assuming that all CODES keywords (omitted from the examples) have the same values as their corresponding VALUES counterparts. [↑](#footnote-ref-3)