Scanner Data; Optimal Preservation Policy for Identifiable Datasets

Meeting of the Group of Experts on Consumer Price Indices
Geneva, Switzerland, 26 - 28 May 2014

Heiðrún Erika Guðmundsdóttir, *Statistics Iceland*, erika.gudmundsdottir@statice.is
Lára Guðlaug Jónasdóttir, *Statistics Iceland*, lara.jonasdottir@statice.is

Statistics Iceland
Borgartúni 21a
IS - Reykjavík 150, Iceland
Tel: +354 528 1200
Fax: +354 528 1299

Abstract

Preservation and deletion of data are among issues to be resolved in the case of large datasets. In the case of scanner data it is important to consider how long it is necessary to preserve the data in full detail. This includes considering all possible uses for scanner data within official statistics, not only in the CPI. Furthermore it is important to contemplate how scanner data can be stored without references or traceability to the chains or stores for a longer period for future research. This implies making decisions about appropriate aggregation and categorisation of the data. This paper discusses these issues with focus on strategies Statistics Iceland intends to employ in this regard.

Key words: Scanner Data, Data Storage, Data Safety, Data Deletion, Confidentiality, Archiving.

JEL: M11: Production Management, M15: IT Management

Rules for Preservation of Scanner Data

Statistics Iceland is preparing a collection of scanner data for measuring prices of food, beverages and other non-durables sold in grocery stores. In negotiations with grocery chains mutual agreement was reached on optimal data structure. Addressing the issue early on and including grocery chains in the decision making was thought to be beneficial for all parties. The grocery chains have been active participants and have given valuable input in the process. It was evident early on that the main concern of the chains was the treatment of confidentiality and the preservation of their data. In short it was considered important that data would be deleted without unnecessary delay after utilisation. This viewpoint is also in coherence with good practice as well as general rules for preservation of data.

Statistics Iceland therefore called upon a few experts and advisors in European NSIs and asked the question: "How long is it necessary for an NSI to keep raw scanner data?" and got some valuable answers. Most of them referred to general rules about the treatment of raw data in their countries but the rules did not specifically address the treatment of scanner data. The answers ranged from two or three years and up to ten years. However there was a general agreement that there would hardly be any need to preserve raw scanner data for as long as the upper limits indicated. Moreover; it would be convenient to keep the data for two or three years. With this advice Statistics Iceland set off to determine a rule to follow in preserving scanner data.

Despite firms' data being well protected by law and good practice it was evident that the chains were concerned that delivering scanner data would be risky, particularly due to the massive detail in breakdown both for products and also over time. With constant focus on price measurements in the compilation of price indices it was surprising to learn that the main concern of the chains was about the quantities being delivered rather than the prices since such information in the wrong hands could be used to project their business strategies. Statistics Iceland consequently decided to analyse thoroughly the need for itemised scanner data preservation as a response to the chains' concern. The focus of this paper will be to discuss the outcome of the process and related security matters that need to be taken into consideration.

Collection of Scanner Data

The objective of negotiations with the chains was to insure a continuous and secure delivery of data, both for the calculation of the price indices and also with regard to confidentiality of data for the chains. When it comes to relying on scanner data instead of price collectors going to the stores, the index production becomes more vulnerable to non-response or difficulties in data delivery. This is both due to the more complex production process and also the short timeframe of the production which constricts the reaction possibilities to non-responses. Some measures were taken in order to counterbalance foreseeable difficulties. A fundament requisition was to make the collection process as automated as possible on behalf of all parties. Two approaches were discussed and agreed to. The first could be entitled the *pull* approach, where the chain establishes a web service that an automatic client from Statistics Iceland logs on to and pulls data in a secure manner. Generally this should occur once every 24 hours and data for one calendar day would be pulled into Statistics Iceland's input databases. If anything were to happen that could prevent the data collection on Statistics Iceland's side the chains are supposed to have readily available the most recent data for up to two months. Furthermore; if the data flow was somehow prevented on the chain side, every necessary measure were to be taken in order to resume data delivery. The second could be addressed the *push* approach, where the chain uploads data to a web service established by Statistics Iceland in a secure manner. Generally this should occur daily as in the other case. If anything were to happen to the delivery process both parties would do everything in their power to resume data delivery and the chain would send up to two months of data if the problem continues for more than a day. In both cases data was expected to be available to Statistics Iceland within two or three weekdays from an end of a business day and for all business days in a year.

Breakdown of Data

The next thing to solve was to determine the variables for the chains to return and how to process them securely. In short, scanner data contains accumulated values and quantities for all sales broken down to individual goods in one day. Data is collected for all days in the year as was mentioned earlier and stored accordingly. From arrival the data will be piped through three main phases. The first could be named *Input*. This is a phase where the data enter in their raw form and undergo their first validation. The second phase could be called

Production as there the data becomes accessible to experts to produce various price statistics. The third and final phase is an *Archive* where the data has been aggregated and links to their origins obliterated. Each phase has its purpose and hence its characteristics will be discussed further in the following sections.

Input phase

The *Input* phase can consist of one or more databases that have the purpose to receive raw data, check for errors and validate the data. At arrival the data undergo integrity examination. This may include checks for number of lines, turnover or data types. All dealings with data in this early stage are handled by automatic procedures. Access to the data is at this point limited to essential IT-personnel and all entries or deletions are automatically logged and managed through stored procedures. Figure 1 lists the variables of data that are stored in the input phase.

Input / Staging

- Entry of raw scanner data

Raw Scanner Data
Chain Identification Store/Outlet Identification Date (Day-Month-Year) Product Number * Product Description * Product Unit of Sale * Product Category Identification * Product Category Description * VAT Level Value Quantity

* In-store specific identifications and descriptions

Figure 1. An overview of variables stored in the input phase. The data are stored in full breakdown and have to be treated with extreme care. Entries or deletions can only be processed through stored procedures. Every action with data in this stage is logged and traceable to a user.

The variables that are stored in this phase are directly linked to individual data providers and are therefore highly sensitive. Store related identification is used to tag the price measures to their origin. Product numbers, as well as description, come directly from the stores. The product numbers are codes that can be EAN numbers, PLU numbers or other types of

barcodes that the chains guaranty to be one-to-one in characteristics. They could even be all of the above. The product unit of sale is the unit that the store uses for charging the product. Where a product is sold in bundles, for example two rolls of bread in a pack, the unit could be “two-pack” or just “piece” all dependent on the store that sells the product.

Statistics Iceland has asked the chains for their own product categories in order to support coding on data arrival, referring to both the automatic coding and the manual coding that follows. The chains had some scruples with this at first given that classifications can be varying between stores; however Statistics Iceland believes that it can speed up the coding process on new product arrivals once the classifications have been studied and programmed into the coding procedures. The chains claim that their own classifications are slowly changing dimensions. It also supports going forward that Iceland is a small country with few retailers predominating the market, where the number of classifications is manageable even though they were as many as the stores. The VAT level needs to be collected for the constant tax rate CPI since it varies across the goods sold in the grocery stores.

Finally there are the aggregated values and quantities of the goods that have been sold on the given day recorded by the date label.

Production phase

When the newly arrived data has passed examination and validation the measures are enriched by coding and transferred into the next phase; the *Production* phase. The production phase is one or more databases that have the purpose to be a workplace for price experts compiling the CPI and other price statistics. The production databases contain detailed data with price measurements that can be directly linked to its origin with the data providers. Access to the data at this point is limited to price experts and essential IT-personnel.

The variables that are stored in the production database are similar to the variables in the input phase with the addition of an assigned product number, link to COICOP or other classifications and the calculated unit value that is derived from the measured value divided by the measured quantity. The assigned product number is a code that is uniquely assigned by Statistics Iceland to the particular good in order to ensure continuity in case the product numbers assigned by the stores are changed. This does also allow the likely possibility that

the same good that is sold in different stores receives the same assigned product number by Statistics Iceland even though it arrives with different product numbers from the stores. The COICOP classification as we know it has developed over time. Statistics Iceland uses its own version of COICOP that is the same as in Europe for the first 4 digits. By allowing alternative classifications the database can keep score of other classification as well as older versions of classifications. These classifications do not necessarily have to be consumption oriented. One example could be related to brand level where each good could be attributed qualities such as *well-known brand* or *brand less*. Figure 2 lists the variables of data that are stored in the production phase.

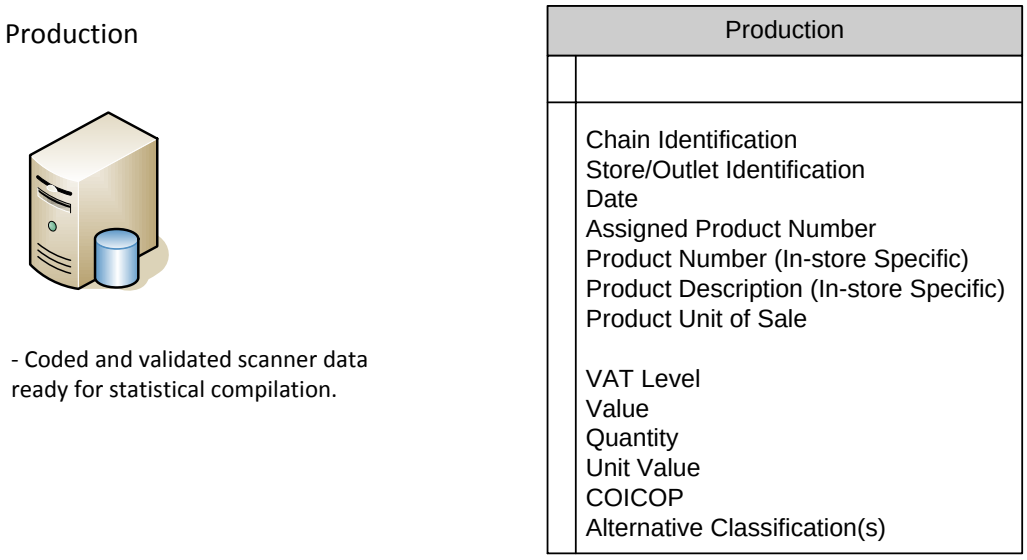

Figure 2. An overview of variables stored in the production phase. The variables are an enriched version of the variables in the input stage. Access to the data is limited to price experts and essential IT-personnel.

The variables that are stored in the production database are highly confidential as before. They lay the foundation for all the compilation that takes place in order to produce price statistics. When a compilation completes, final results are prepared for dissemination.

The Archive

After dissemination final results along with aggregated data are transferred into the next and final phase; the *Archive*. The archive is one or more databases that have the purpose to store data for future reference or research. Access to the archive is a *read-only* access and available only to price experts and essential IT-personnel. The archive can function as a

warehouse in the sense that extracts from the archive can be made available to other groups of experts that use results from price statistics as inputs in their production, e.g. national accounts, etc. The archive is however different from the other phases by containing no references or links to the origin of the data. Figure 3 lists the variables of data that are stored in the archive.

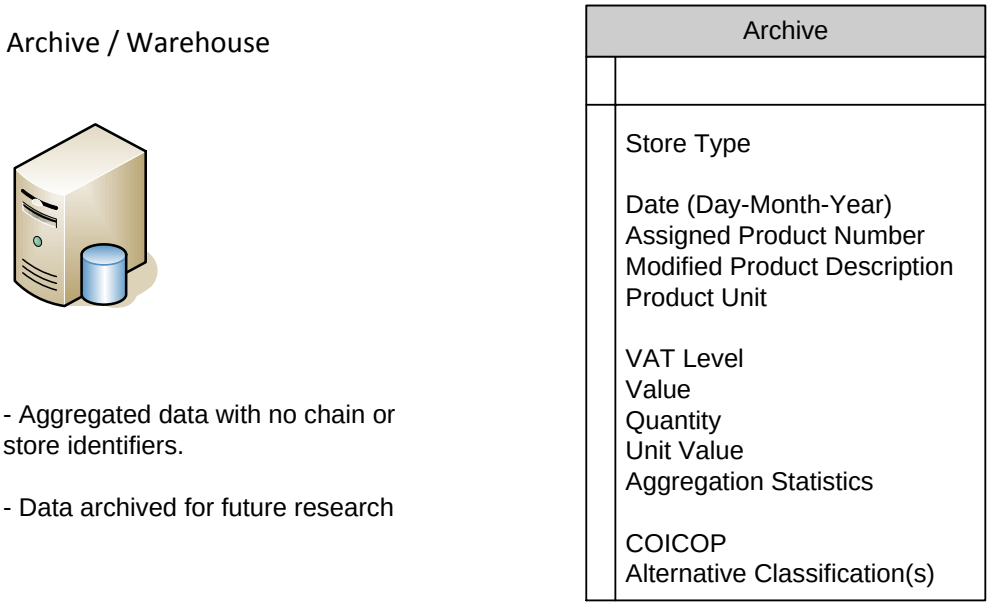

Figure 3. An overview of variables stored in the archive. The variables are an aggregated version of the variables in the production stage. They contain no link to the origin of the data and aggregated values cannot be traced to a specific data provider.

The archive is a platform for storing data for an indefinite time without having to worry about confidentiality being compromised. In the long-term perspective the aggregated data becomes the focus. The original raw data or links to their providers are no longer relevant. Therefore there is a new structure of variables in this final phase. As before the archive stores values, quantities and the derived unit value; however in this case the measures are aggregates or averages for specific store types. Store types can be e.g. *supermarkets*, *discount stores*, *all hour stores* or any other type that is relevant in the market. A different setup of store types could be *all stores in the capital area* and *all stores outside the capital area*. The summarised measurements can be stored with a link to a day or a month or another time period. They can also be stored with some statistics such as a maximum or a minimum value. An assigned product number will ensure consistency and link the product to classifications that can alter with time. Furthermore it could be interesting to store

aggregations to product groups instead of individual products and then a new assigned product number could be issued for the group.

There is one more variable worth mentioning and that is the modified product description. If the archive relies on the product descriptions originally provided by the stores it has to be guaranteed that all indication or symbols relating to shop own brands (s.o.b.) are removed. This can be done either by rewriting the product description or by substituting the brand link with a string like “sob”.

An archive is not limited to storing only one type of aggregation from the production database. It can store aggregation from the complete set of data in production. It can also store aggregations drawn only from the set that was used in the CPI. The possibilities are many. The only restriction is that the archive has to be completely exempt from links to individual data providers and any traceability that could lead directly or indirectly to them.

Preservation Requirements for Use in CPI and HICP

The bulk of grocery store scanner data will be used for month-to-month price comparisons. Even with a fixed basket approach the basked would in most cases be updated annually. Regular production would therefore, in the strictest sense only call for 13 months of data to at hand. There are however more scenes that need to be viewed. Irregularities arise with missing prices or prices of seasonal goods. Moreover; they can arise with new products and products that appear in the market for only a short while and then disappear. Such products have induced many countries to restrict their use of scanner data to products that have been in the sample for a predefined period before allowing it to enter compilation. The policy for how long the predefined period should be varies between countries. The countries also take products into account in their decisions.

> A product (EAN-code) will have to have a life span of at least two consecutive months to be eligible for price index calculations for a monthly index. In many cases countries using scanner data have set stricter criterions of a products lifespan before it is included in index calculations. For example, Denmark has minimum standards for how long the product has to have been available in the data and looks at the share of each product for its respective product group when selecting products for their calculations. In most cases a lifespan of at least 12 months is required but sometimes less (Gustafson, 2013).

Statistics Iceland has not decided upon a policy toward a minimum pre-lifespan period, but will look towards practices in other countries in addition to learning from the Icelandic scanner data. All the same Statistics Iceland does not view the pre-lifespan of products to be restricting to the preservation policy since an existence of a product in a time period can be accounted for even if the related measurements have been deleted.

Irregularities caused by missing prices or seasonal products do however have to be addressed. The problem with missing prices is quite relevant in the case of a fixed basket approach where the basket is small or where each category has only a few measured goods. However when a good has been missing and inquiries report that it will not be returning the good is replaced by a more suitable one. Missing prices for food products do not cause extensive problems in the Icelandic CPI. The market relies on imports of seasonal products like fruits and vegetables for the greatest part of the year thus most products are available all year-round. Therefore Statistics Iceland puts missing prices aside when considering the necessary lifespan of the data. Even if they did propose a greater difficulty the more extensive basket of goods that is available with scanner data should also lessen the effect.

The case of seasonal products may be more relevant. Many traditional seasons have the regular twelve month recurrence whereas others have a more varying pattern. The seasons that have the most effect on product variety in Icelandic grocery stores are summer, winter, Christmas and Easter. The summer and winter seasons may be considered to spread over a few months each even though their climax may vary and consequently affect consumers buying pattern.[1] Christmas has a constant twelve month recurrence rate since, however the recurrence rate for Easter can be from about 11 to 13 months.

Preservation proposition

Seasonal products usually enter the market before the high season and can stay on the market for a while afterwards even though the sales drop. Taking that into account as well as the varying recurrence pattern of the seasons, Statistics Iceland suggests that a period of 18 months should be sufficient for the raw scanner data to be fully exploited in the CPI and the HICP and hence for the preservation of raw scanner data. All price comparisons should

[1] The summer e.g. is considered to be June, July and August. The weather, that can be very interchangeable, plays a big role in the decisions of the general Icelandic consumer, whether to buy summer products or not. If June is a good month (calm and warm) this for example greatly influences the sales of meat for barbequing. The next year it may happen that the good month is August.

have been made within that time frame. Statistics Iceland also believes that 18 months of data should be more than enough for selecting which products should enter compilation of the indices.

Other Uses of Scanner Data

Statistics Iceland's main use for scanner data is compiling the CPI and HICP. Both indices are calculated from the same dataset and hence the same rules apply for processing the data in both cases. Statistics Iceland hopes to be able to use scanner data for PPP consumer surveys in the future but at the moment the focus in on implementing the use of scanner data in price indices. Many countries have already started using scanner data for their PPP surveys. The longest possible process for a PPP survey as it is managed by Eurostat now is about 20 months. This counts from the very beginning when countries receive preview questions in preparation of a survey until the last validation round has been completed. The first phase of a PPP survey is the preview where the PPP team scans goods for pricing in the survey. If the supply of scanner data is up and running the PPP team will have up to 18 months of detailed data to research for good items. The next phase is the presurvey where the search continues in more detail. The actual survey where the price rating starts takes place approximately a year later after the initiation of the preview. This means that from the time of the price rating until the end of the last validation round there can be no more than 8 months. Hence the proposed preservation time is supposed to be sufficient for the PPP surveys if they were to rely on the scanner data.

Conclusion

After considering the particulars of utilisation and preservation of scanner data the outcome is that it is sufficient to keep scanner data in full breakdown for 18 months for use in the CPI, HICP and PPP. For use in research that relies on longer time series the raw data needs to be aggregated and archived without all links to stores or chains. This entails that automatic processes have to be implemented to delete raw scanner data from the input and the production databases before 18 months have passed since their effective date. Before this deletion happens aggregations need to be completed and transferred into the archive. Figure 4 maps the data flow and relative processes.

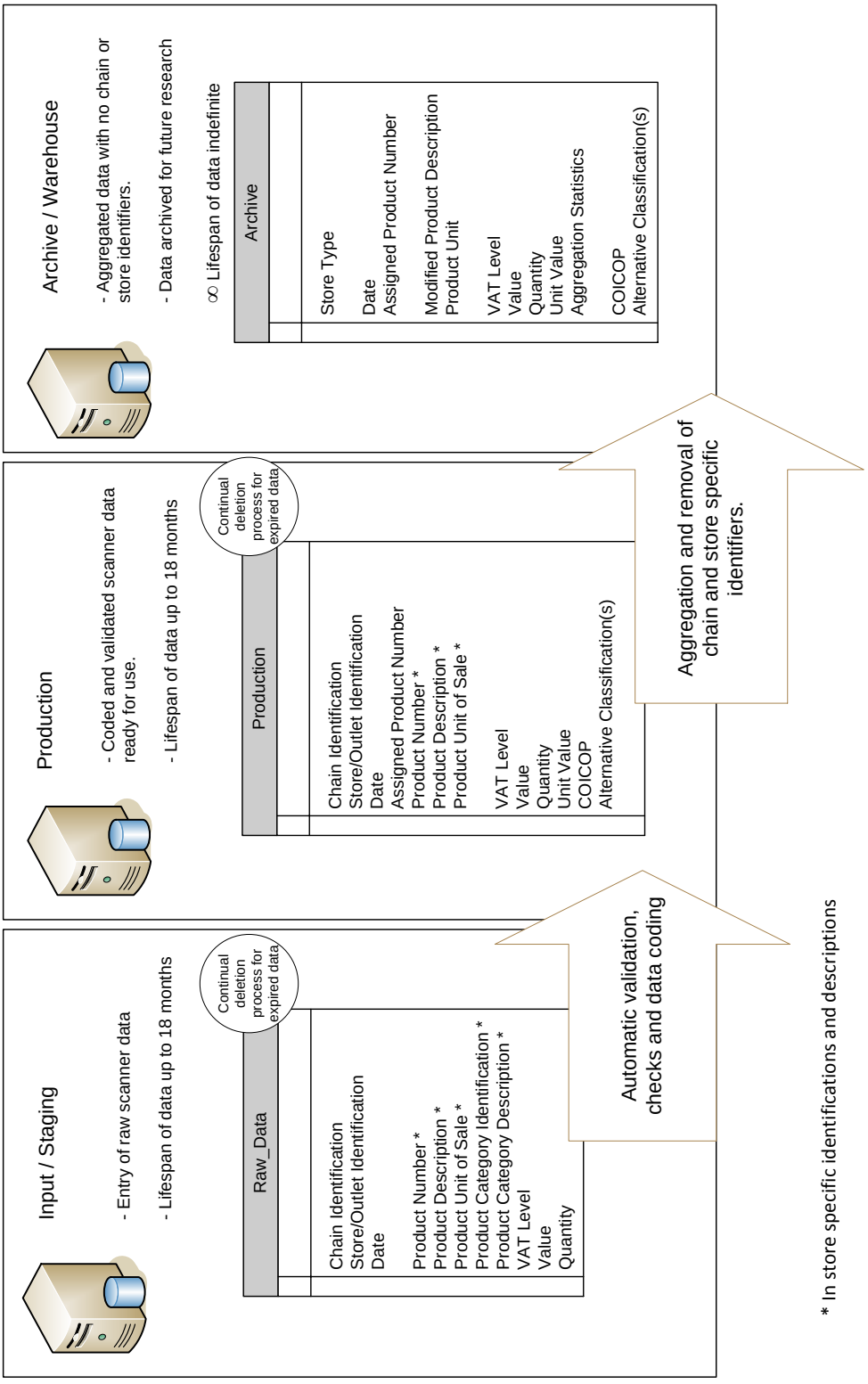

Figure 4 The data flow of scanner data from the arrival in Statistics Iceland until deletion. Automatic processes manage validation and transfer of data from input to production and from production to archive. In addition automatic procedures maintain continual deletion of expiring data in the input and production databases.

As can be seen in the figure automatic processes manage transfer of data between phases as well as deletions of expired data. Access to each phase is limited to essential personnel. Data can exist in input and production databases for up to 18 months until they expire and are deleted. Aggregated data in the archive has no expiration date and can be stored indefinitely, as all conditions for confidentiality are fulfilled. As such the archive can be a valuable source for future research in the field of price statistics.

Statistics Iceland would like to report that even though the proposed 18 months are considered a sufficient time to keep scanner data in full detail for processing price statistics then the Icelandic grocery chains have agreed to let the institution keep their scanner data for 36 months to begin with while the implementation is being prepared. That means that after 36 months of receiving scanner data Statistics Iceland has to delete the oldest 18 months of the raw data and initiate the continual deletion of data that become 18 months old.

References

Gustafson, N. (2013). *Drawing a Sample from Scanner Data to use in the Danish CPI*. Paper presented at poster session at the 13th Ottawa Group Meeting, Copenhagen, Denmark, May 2013.