26th Meeting of the Wiesbaden Group on Business Registers
- Neuchâtel, 24 – 27 September 2018

Steen Eiberg Jørgensen
Statistics Denmark
Session No. 5 – New data sources

Big Data – New data sources

Paper

Over the last ten years, Statistics Denmark (SD) has carried through a number of initiatives for the purpose of digitising the data collection for business surveys, and the data transmission from the enterprises to SD is now essentially digitised.

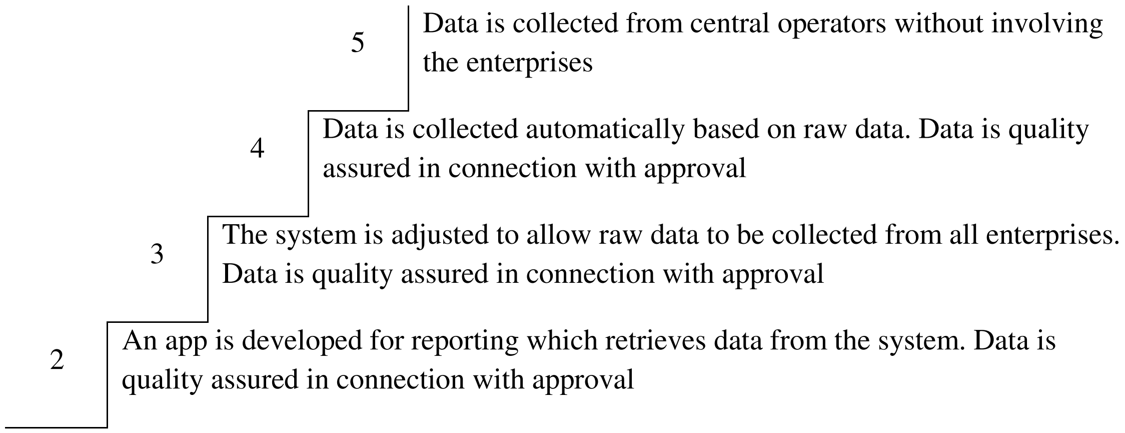
In the course of this digitisation process, the technological development has resulted in even more work processes and systems becoming digital. In many fields, this means that data is born in digital form, and handled and stored digitally. This applies to e.g. ordering on an enterprise website, which is then invoiced digitally and subsequently stored automatically in a cloud-based accounting system. The digitisation implies that data is increasingly shared between systems and makes workflows smoother and faster. In addition, there is a tendency for enterprises to share data if it can contribute to reduce administrative tasks.

When systems become online-based, it further results in an alignment of the application of the systems by the enterprises. Customised use is not possible to the same extent and, consequently, data is registered and stored in a consistent way. The consistent structure of data facilitates the use of data for statistical data collection from many enterprises in preference to small samples and it eliminates non-response error.

As a result, the technological development makes it is possible to extend the digitisation process into the enterprises, so that data is collected directly from the enterprise systems. In this way, a change has happened by which digitisation transitions to automation. The digitisation into the enterprises and the automation of the work processes can be implemented at various levels.

Figure 1

Various levels of automation



5 Data is collected from central operators without involving the enterprises

4 Data is collected automatically based on raw data. Data is quality assured in connection with approval

3 The system is adjusted to allow raw data to be collected from all enterprises. Data is quality assured in connection with approval

2 An app is developed for reporting which retrieves data from the system. Data is quality assured in connection with approval

1 The enterprise can extract and submit a standard file from the system. Raw data is collected without an actual approval process.

0 The situation today — the enterprise can report digitally via a form at Virk.dk with quality assurance of data

What is the current situation in Denmark?

The Danish Parliament has decided that data must be shared to the extent possible. For this purpose, they have established the Agency for Data Supply and Efficiency, which for the past 12 years has helped to promote digital solutions and often common solutions in the public sector. Moreover, the aim is to be able to share reports/registers with the private sector, unless they contain sensitive information. In this way, enterprises can report to SKAT (the Danish tax authorities), the Danish Working Environment Authority, SD, etc. via the common public portal for reporting - Virk.dk. The agency is also seeking to ensure the sharing of data to minimise instances where central and local governments are requesting the same information twice.

Today, the digitisation of the data collection in Statistics Denmark is close to 100 per cent. Reporting by paper is only possible for enterprises that have been granted an exemption or that report for non-statutory statistics.

SD expects to receive 450,000 reports distributed on approximately 120 sets of statistics in 2018.

There are four reporting solutions:

Upload – manual data entry –
applied for small sets of statistics where the enterprise enters data in a spreadsheet, with limited validation.

Digital forms – Manual data entry with online validation –
is the solution in which SD has been the most engaged throughout the last 10 years. The forms are very similar to the questionnaires on paper that we used previously, and each set of statistics has its own form. However, the solutions incorporate ever more troubleshooting to reduce the need to repeatedly contact the enterprises.

System to system solutions – Reusing data –
Applied on a few sets of statistics where the enterprises give us access to retrieve data from their own systems and to send them without further processing to SD, e.g. the hotel bookings systems, payroll information, external trade, sale of cinema tickets.

APP – Automatically generated entry of certain data
Solution applied for the statisticTransport of goods by lorry , where the driver is often responsible for the reporting.

Digitisation trends

We see more and more online systems that standardise the enterprises, and we also see that the system suppliers are few and large. Today, primarily small enterprises use these common systems where data are not stored within the enterprise itself but in a cloud solution.

In small and especially in new enterprises, data is born in digital form. When you create an enterprise, you do so digitally by means of your personal digital_ID. It takes five minutes to create an enterprise, after which you can retrieve standard articles of association. After that, you can create an account as a customer of an accounting firm and a payroll service agency then you are up and running 100 per cent digitally.

What new and small enterprises seem to have in common is that

- Operating a business should be easy
- Administrative tasks should be minimised
- Data sharing is acceptable if it makes life easier

Enterprises as well as data collectors and users want to cooperate with SD based on a wish to gain more knowledge on data, make reporting easier, improve quality, and they have confidence in SD as a business partner.

The biggest challenge is often to have the imagination to find the data collectors.

Various levels of automation

We currently experience five levels of automation in the reporting process as illustrated in figure 1.

Level 0 is where we are today with a few exceptions. At this level, we ask the enterprise to report digitally and the workload reduction has primarily taken place at SD.

Level 1 is a manual process where the reporting is based on raw data which is transferred manually from one of the enterprise's systems to SD. The submitted data can be subjected to a manual quality check, but this is not a condition of submitting the report.

Levels 2 and 3 are semi-automatic processes during which data is transferred from one of the enterprise's systems to SD. The reporting process must be activated manually and data is presented to the reporting office for approval before it is submitted. If necessary, any missing data is entered manually. There is an option to implement troubleshooting rules in this approval process.

Level 4 is a fully automatic process during which a component is made available that automatically generates the report based on data in various systems in the enterprise. Missing data is automatically interpolated in the reporting process, e.g. by means of machine learning. Data is presented and validated by the reporting office before it is submitted. There is an option to implement troubleshooting in this process, and subsequently the reporting office approves the data.

Level 5 is a fully automatic process that collects data from central operators in the market. The process ensures that it is not necessary to ask the individual enterprise to report. Instead, the reporting is performed by central operators on behalf of the enterprises that would otherwise have had to make the reporting. The process opens up the possibility of entering into a digital partnership, so that the central operator not only submits data for the enterprises that are subject to reporting duty, but instead submits data for all enterprises with which the operator is cooperating. As a result, SD will not only receive sample data, instead we will receive complete data sets for the relevant market segment.

Possible action areas:

At present, we have identified a number of action areas which are all relevant to develop further in the years ahead:

- 120,000 small and medium-sized enterprises use the same online-based accounting system. Altogether, the five dominant enterprises in the market have 200,000 customers of the total potential of 320,000 enterprises. The online-based accounting systems are progressing according to the customer needs, and there is reason to expect that enterprises using an online tool today will continue to do so. We are in dialogue with one of these suppliers and they are very positive towards developing a solution that allows enterprises to sign up for submitting data to Statistics Denmark at automation level 4 or 5. Against this background, we also believe that the other suppliers are going to participate, since this is a competitive parameter.
    - An example of use could be turnover figures which can be collected on a monthly basis.

- More or less all the major farmers are using the product Cloudfarms. Cloudfarms has agreed to enable reporting to SD via Cloudfarms, corresponding to automation level 3. Correspondingly, Cloudfarms is interested in providing data for all pig farmers to SD on a weekly basis, corresponding to automation level 5.
    - The pig population: reports for this are submitted on a quarterly basis by 2,300 pig farmers. A third of all pigs are reported by 120 major pig farmers using the online system Cloudfarms as a feeding management system.

- In terms of the statistic Transport of goods by lorry , there are also major suppliers of fleet management tools. This can contribute data to a wider extent on the journeys of lorries for transport companies at level 5. The collection of data from the fleet management systems means that SD will receive complete data sets from the operators in question and not just sample data for individual lorries. Data about the goods they transported will still be missing. For this purpose, we will probably still need to develop system to system solutions, or await digital consignment notes. Alternatively, ease the response burden for lorries driving for one type of company, e.g. convenience goods. For the major transport companies, all data in the major transport companies' fleet management systems is available with a few exceptions. It is anticipated that data can be collected directly from the fleet management systems (only in individual file formats) corresponding to levels 1, 2 or 3, depending on the willingness of trade associations and system providers to cooperate and pay.
    - Information about Transport of goods by lorry is reported on a weekly basis by 170 selected lorries.

- A current project about automatic business reporting will enable generation of reports for the accounts statistics directly from the enterprise accounting systems and other connected systems, e.g. the inventory management system. It is expected that information which cannot be found directly in the system, can be interpolated by means of machine learning (or manual entry), after which the information can be approved and submitted by the reporting offices. This will automate the reporting for the accounts statistics for the 8,000 annual reporting offices, corresponding to level 4. Via the Danish

Business Authority, SD receives accounting information for the enterprises not included in the sample for the accounts statistics, corresponding to level 5.

  - The accounts statistics has 8,000 annual reporting offices, as well as key figures for the remaining population of enterprises at a legal level.

- There are two central reporting offices for statistics on felling in forest and plantation and forest area; HedeDanmark (a service and trading company within the green area) and Skovdyrkerforeningen (an association of silviculturists). Experience has shown that the information to be reported exists in digital format with these two central operators – for the enterprises subject to a reporting obligation as well as for others. A digital partnership in this field will enable automation of major parts of this survey. This solution will correspond to level 5.
  - Of the 1,100 reporting offices, two companies account for a major part of the reporting on felling and forest areas.

- Information on harvest is widely available in cloud-based databases that collect the information directly from combine harvesters applied for the harvesting. Cooperation with the suppliers of the combine harvesters will allow us to obtain an automation corresponding to level 5.
  - Harvest of grain, rape and pulse: information about harvest is reported annually by 2,700 farms.

Clarification issues – challenges

A number of issues are expected to need clarification in respect to collection of reports directly from new data sources.

In terms of collecting data from central operators, the legal authority to do so must be clarified. Under the authority of the current Act on Statistics Denmark, we can legitimately collect information from enterprises concerning the enterprise itself. The act does not currently authorise us to collect information from central operators involving data concerning other enterprises. This means that collection of data from central operators must be based initially on voluntary digital partnerships.

In terms of automatic reporting from the enterprises, it needs to be clarified to what extent the system suppliers are willing to co-finance the adaptations to facilitate the automation. Automatic reporting can be a competitive parameter, especially for new online-based systems which are trying to gain a foothold in the established market. Our first dialogues with the involved parties have been positive and the enterprises/associations do see an advantage in making solutions available to SD.

In this connection, it is also relevant to clarify the extent of diversity in the systems applied by the enterprises. The migration from offline to online systems means that the enterprises more so than previously accept that they have to use standardised charts of accounts etc. However, this may be a challenge, if there is major diversity in the file formats submitted, if a manual process such as level 1 is concerned. If data is submitted via an app or as system-to-system reporting, it will facilitate alignment of the file format.

Finally, the role of the trade associations in respect to automatic reporting and digital partnerships must be clarified. It is in the interest of the trade associations to reduce the reporting workload for their corporate members, and for this reason they may want to co-finance functionality that allows data to be collected via an app or a system-to-system solution.

If they enter into a voluntary digital partnership, it will impact enterprises subject to a reporting duty so that they no longer need to approve the information that is reported. This is why it must be clarified whether the enterprises are required to authorise that the information is submitted by the central operator.

In addition, it must be clarified whether the central operator is allowed to submit the information to SD for enterprises that are not subject to a reporting duty. This probably requires that it becomes part of the business concept for the central operator to regularly submit data to SD.

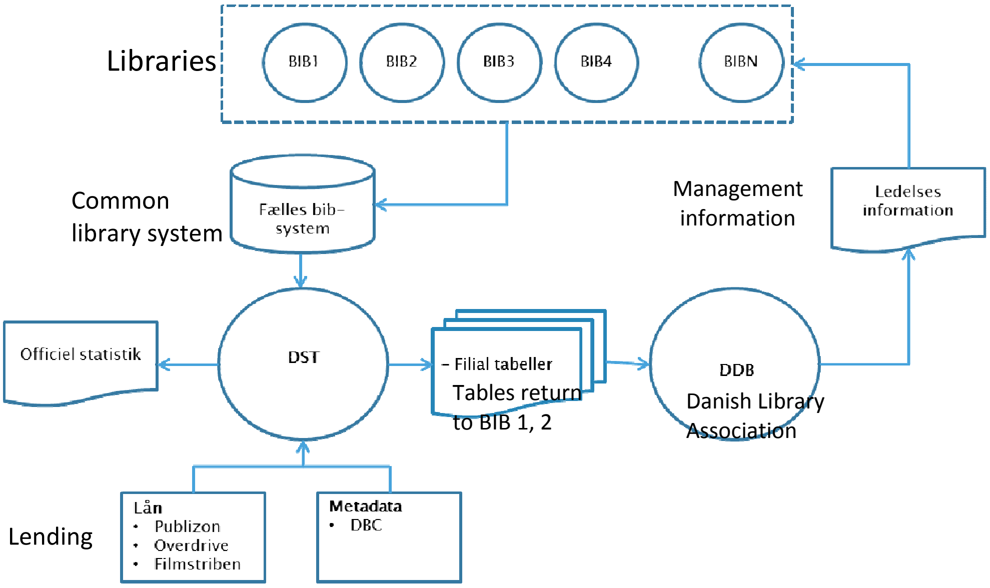
However, the central operators' permission to submit data is not expected to be a relevant issue, if the central operator owns the data to be reported. Accordingly, the ownership of data is also an area that needs to be clarified.

The public sphere

During the past few years, Statistics Denmark has assumed the responsibility of developing the cultural statistics in Denmark. Among these are the statistics of the lending of books, which we collect from all Danish libraries today.

The future data collection will be digital using SD as a data business partner and in the long term as a data processor. The system is designed as follows:

Figure 2



This means that SD combines forces with the municipalities to develop a new IT system that allows libraries to also benefit from reports to SD.

Role of the Central Business Register

The Central Business Register will still play an important role in terms of being in control of the units, follow units over time, industry classification, geography, compilation, population etc.

It will be possible to use some of the new details for industry classification as well as for detecting mergers and acquisitions via the annual reports from the enterprises.

We focus our energy

To make the most of our efforts, SD will give priority to solutions that incorporate agreements with data processors. The suppliers seem favourable towards this. Furthermore, it will be the least costly solution to implement, and it will reduce the workload on the enterprises the most.