Automated data collection and reuse of concepts in order to minimise the burden

Anna-Greta Erikson, Johan Erikson, Cecilia Hertzman[1]

Abstract

For statistical offices and other government agencies, there are requirements to simplify reporting for enterprises. For a number of years there has been a move from paper questionnaires to online questionnaires and e-services. Now is the time to take the next step in the automation of the data provision. In the simpler form, data is imported via a file from the business system to the questionnaire, and in the even more automated form, fully automated file transfers where no person need to press "send" (machine-to-machine).

Statistics Sweden has for some years offered the possibility to import data from accounting systems into the online questionnaire. This has proven very successful in terms of both response rate and response burden, while the quality of the statistics increased. Working with these more automated modes of data collection requires both changes in the technological infrastructure, and the way we organise the requests of data to the enterprises. Data requests need to be built in a slightly different way and adapted for what is available within the business system.

In order for these changes to be as effective as possible from a data providers perspective there is also a need for collaboration with other authorities, requesting similar information. Ensuring the use of concepts, that the same concept has the same definition and meaning in different reports is a requirement for effective machine-to-machine solutions.

Together with the Tax Authority and the Swedish Companies Registration Office, Statistics Sweden has launched a project with the aim to reuse financial information. The format to be used is XBRL (eXtensible Business Reporting Language) and the intention is that an enterprise will be able to produce the information once, and reuse information to several authorities. The paper discusses the experiences of this collaboration and lessons learned so far.

Key Words: Coordination, Data collection, Digitalisation, XBRL,

1 Background

A statistical agency faces several major challenges in its business, such as: changes in the outside world and new user requirements, increased non response, demands for efficiency, and demands to reduce response burden and to face the technological development. In Strategy 2020, Statistics Sweden has as one of the overall goals to “meet

[1] Anna-Greta Erikson, Statistics Sweden, SE-701 89 Örebro, Sweden, email: annagreta.erikson@scb.se. Johan Erikson, email: johan.erikson@scb.se. Cecilia Hertzman, email: cecilia.hertzman@scb.se

the needs of users for statistics of high quality". Another overall goal is to "make it easy to provide the right information". To reach the goals one prioritised way is to work for a uniform and coordinated statistical system with consistent data. Another stated way is to use uniform, standardised and efficient processes, tools and methods: i.e. to facilitate the response process and to secure input data, Statistics Sweden should offer several ways to provide information, adapted to changing needs as well as to new technologies. One way to go here is to use the on-going digitalisation in society to offer more automated ways to provide data.

The idea is that by working to coordinate statistics to improve the quality of statistics the burden is also reduced simultaneously. By coordinating the contents a better coherence and comparability for users is obtained. When the content is coordinated the concepts used in data collection are coherent and clear for respondents and therefore the reliability of statistics is also affected positively. Coordination of the contents of economic statistics and the ambition to reduce response burden for businesses interact with each other.

2 Digitalisation in data collection

Digitalisation in itself is not new for businesses. They have had administrative systems for book-keeping, wages and salaries, sales and other operations for a long time. However, data provision to authorities for a long time relied only on filling in paper forms. In the example in Figure 1, we can follow an inefficient flow where the same data is digitised twice, in the "analogue loop of stupidity". The data is digitised first in the enterprise's own process. Since there is no digital service available for providing data to an authority, the information to be reported is printed out and filled into a paper questionnaire which is delivered by traditional mail. The report arrives to the receiving authority where the incoming documents are scanned or manually registered, or in some cases only stored as images (e.g. annual reports sent to the Companies registration office). In the latter case, several users of the information in the documents (i.e. credit agencies) then purchase the scanned images and standardise and digitise the numerical data for their needs, this actually is the second time the same information is digitised.

Figure 1: The analogue loop of stupidity (Source: Swedish Companies Registration Office)

Over the last 10 years, data provision has been moving gradually to the internet, but in truth this digitalisation process is only half-baked, even if the questionnaire can be intelligent and offer help, routing, quality checks and other aid in filling in the data. It is still reliant on a person filling in a questionnaire, albeit using a computer instead of pen and paper, so only the second part of digitising the data is eliminated. In order for digitalisation to have a full break-through in the communication between businesses and authorities, it is necessary to take one step further, and allow businesses to directly transfer data from their administrative systems to the authorities requesting information. This normally goes by the name "machine-to-machine" data collection since it is set up only once and then over time runs automatically without the involvement of a physical person. Such a process has the potential of dramatically reducing response burden on businesses, and such solutions are also more and more requested by the businesses. Looking at the digitalisation process over time, it can be described as a continuous development where data provision is moving in stages from traditional paper forms to digital solutions, see figure 2.

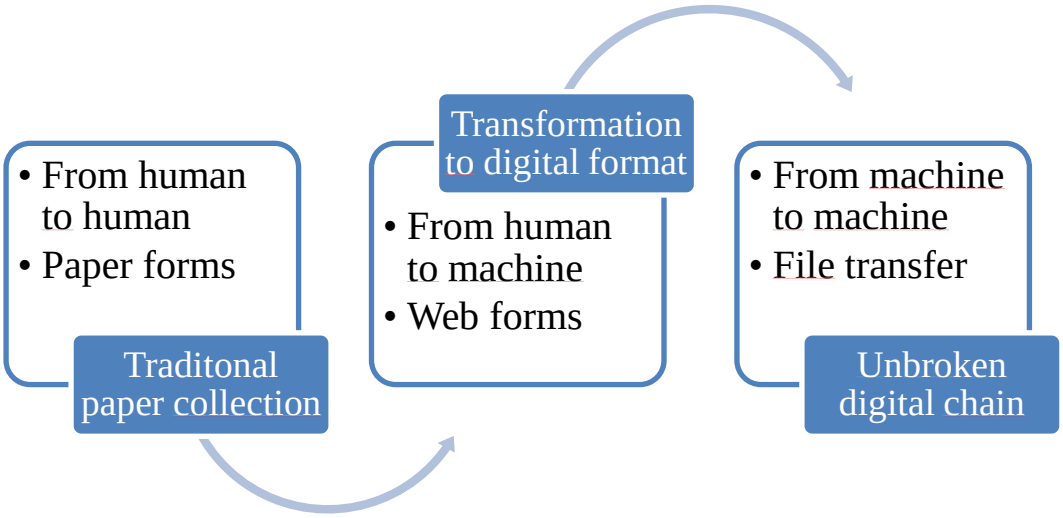

Figure 2: Digitalisation in data collection

2.1 A step-wise approach to digitalisation

For reporting of financial information, in the broadest sense, to governmental agencies, there is need to provide different solutions in different areas; some information requested is more or less directly accessible in the enterprise's administrative systems, while other requested information is not possible to simply access in an existing administrative system. In the first case a file transfer or a machine-to-machine solution may be suitable, while in the latter case there may be a need to enter data manually to fulfil the requests, and the web form suits such requests well. When data requests are a mixture of the both cases, a semi-automated solution could be the appropriate one to offer. There is also the possibility to have semi-automated solutions with a large part of automation where respondents extract a file from their systems and send them in without further treatment. For data requests that are both recurring and tailored to what's available in business systems, it is possible to take the digitalisation even further, building completely

automated solutions. So providing a floating scale of digitalised solutions, we can divide them into the following categories:

- Half-digitalised solution: Web questionnaires. This will not be covered further in this paper.
- Semi-automated solution: Some data provided from systems, but manually completed.
- Semi-automated solution with a high degree of digitalisation: All data in a request is extracted to a file, which is sent manually to an authority.
- Fully automated solution: Machine-to-machine solution, where data is sent from a business system to an authority automatically, without any manual work needed. In a fully integrated solution, the data provision is completely integrated into the businesses' own processes, where the necessary data is generated in the business processes and automatically transformed into the requested formats. Integration with business processes is of course possible also in semi-automated solutions, the fully automated solution adds the automatic transfer.

2.2 Semi-automated solution

In the scale above, two levels of semi-automated solutions are described. The first is where only some of the data in a request can be extracted from a business system. One example of such a solution is that for several years Statistics Sweden has been offering respondents in the annual Structural Business Statistics (SBS), the opportunity to import accounting system data in to the web questionnaire, instead of filling in the form manually. It has proven very successful with regards to both response rates and response burden.

The file format used in the solution is the standardised Swedish format, SIE (Standard-Import-Export). The SIE format is an open standard for transferring accounting data between different software produced by different software suppliers. SIE could be used to transfer data between software on the same computer, but also used for sending data between companies, for example between the company and the accountant. In this specific case the transfer is between the company and the statistical office. In this case an already existing business process that generates data is used for an additional purpose, to use that data for sending information to the statistical office. The SIE standard is well spread in the software business in Sweden and it is a de facto standard for transferring accounting data. Another infrastructural component that is crucial for this semi-automated solution is BAS, the leading chart of accounts in Sweden and in practice The Swedish Chart of Accounts. About 95 per cent of the businesses in Sweden use BAS. SIE and BAS are fully compatible.

In the SBS data collection, the respondents upload the standard SIE file with accounting data at BAS-account level and where it is possible to link BAS-accounts to SBS variables, the information is translated into SBS variables and prefilled into the web questionnaire. For some prefilled variables there is a need to do an additional manual split into several other variables. In the questionnaire there are also SBS variables without a link to BAS-account, which must be filled in manually.

The reactions from respondents have been overall positive, Statistics Sweden has gained all time high response rate, lower response burden and moderate development costs. In the last completed survey, 43% of the respondents chose this way of reporting. The

average time for filling in the ordinary questionnaire is about 60 minutes while by using SIE the time has been reduced to 30 minutes.

Of course, this approach can also be used in other cases and using other formats such as text files or excel files as well. Several such cases have been set up at Statistics Sweden the least two years, allowing businesses to import data from a file to prefill parts of a web questionnaire.

The second stage of semi-automated solutions is where all data in a request can be provided by sending a file instead of filling in a web questionnaire. This normally goes under the name of "file transfer" and is used extensively in a number of surveys at Statistics Sweden, especially in public sector surveys. Here, the respondent is given a choice of either filling in the questionnaire or sending a file through the web portal. The reason that this is called semi-automated is that it still requires the respondent to log on to the portal, choose the correct file and upload it. It should be noted though that many respondents appreciate these solutions since they still are in control of which data is sent and when.

2.3 Machine-to-machine

For recurring requests, moving from a semi-automated file transfer to a fully automated machine-to-machine solution can be a desirable solution. This means that a contract is set up between the sender (the enterprise) and the recipient (the authority) on data content, format and transfer scheme, and an automated recurring data transfer through for example file transfer protocols or open APIs is set up.

Statistics Sweden offers machine-to-machine solutions for some surveys. However they are not that well-coordinated and have been set up for each survey separately. In order to allow for machine-to-machine data provision on a larger scale, the processes within the statistical office need to be somewhat adopted. Instead of sending out requests and collecting data, this means more work with setting up and maintaining contracts and technical solutions for automated transfers, and also checking that expected data actually comes in according to the transfer schemes.

2.4 Integration with business processes

To integrate the generating and reporting of statistical information into the ordinary business processes means that creating the data to be reported is an integrated part of an unbroken digital chain within the enterprise; from the moment the business transaction is registered until the aggregated data is reported to the appropriate authority. In an optimal situation, the data reported occurs as a fully integrated part of the normal business process. Transactions are for example recorded in the accounting system and quality assurance is done in this process. When these transactions are aggregated and form part of i.e. a tax return or a statistical inquiry, the originality and quality is guaranteed through the unbroken chain. In an extreme case, data could actually be generated and transferred on a transaction base, but it can also be the case that the data is generated in the business process and transferred as aggregated data on a regular basis through a machine-to-machine solution.

Integration into business processes can be used for both semi-automated solutions and machine-to-machine solutions. In the section above, describing the use of the semi-automated solution (see 2.2) for the Structural Business Statistics, Statistics Sweden makes use of the fact that the normal accounting system can produce an export file with

general ledger data in a common structure. The content of the export file is the same that normally processes in the program where the annual report of the enterprise is compiled. In this semi-automated case the output from the business process is used extensively, and further processing and manual supplement is made where the existing business process cannot provide information at the requested level of details or in a common structure. Data may actually be available, but not in such a standardised way that it can be used for automatic pre-filling of a questionnaire. In the BAS chart of accounts, and therefore also in SIE files based on BAS, some accounts can be used more freely, why the same account number can be used for different things by different enterprises. Extracting that information in a standardised way is much more difficult, but could be possible by for example reading the account tags and use text mining techniques. However, this is not something that Statistics Sweden has tried yet.

3 Coordination for more efficient data collection

In the current situation Statistics Sweden may offer solutions that are quite automated, but not in all surveys, and not coordinated. This means that one function of the enterprise may need to communicate information to Statistics Sweden via several channels; i.e. via traditional web forms, via various file transfer solutions, some semi-automatic and some more or less machine-to-machine but probably to different server addresses.

Thus there is a need for coordination of recommended technical solutions that should be offered to respondents in different kind of surveys. In order to make it possible to further digitise and automate the data provision we, as a statistical authority, also need to coordinate and re-design our data requests. We need for example to adapt our data requests to business definitions to be able to use what is in the business systems and we need to re-structure what is asked for in which survey. Further, of course, also technical solutions need to be rebuilt, but that is not where the greatest effort is needed. The major challenge is in coordinating the content.

3.1 Coordination of concepts and survey content

Many surveys do revise their content occasionally or even frequently, but there is still a large gap between what information and definitions surveys ask for and what is actually available in business systems. Many surveys ask businesses to adapt to our statistical definitions by including and excluding data based on detailed descriptions and definitions. For a digitalised data collection, this needs to be turned around. We need to collect what is actually in business systems, and then transform that information ourselves to statistical concepts.

In order to build efficient solutions, coordination of concepts and definitions between our surveys is required. It will be easier for businesses to understand what is asked for when the same terms are used across various surveys and in reports to several different authorities. Unnecessary variation creates confusion. In this way, the same concepts and definitions can be reused and efficient solutions can be built by software companies or the respondents themselves. It also opens up the opportunity to share information between authorities when alignment of concepts has been taken place and it is clear what is being reported.

This coordination of concepts and the adaption to business language increase the quality and coherence in provided data, but it also raises many issues for the statistical office

such as methodological questions. Relying on business definitions and what is available in business systems will probably mean that different information than today will be collected. We will have to rely more on model assumptions to transfer this collected information into the statistical concepts, sometimes also using our own assumptions rather than to ask the businesses to make estimates or approximations. Changing the survey borderlines between surveys (see below) will also mean that how the sampling design is connected to individual variables may change, so that some variables are collected from fewer enterprises than today and other variables from more enterprises. This will in turn affect our estimation processes. However, it could be worth also asking ourselves which quality the data collected today has if we ask for information that is not really available in businesses.

3.2 Re-design of survey borderlines

Many surveys have a long history and were created for specific purposes of their users. But a fully digitalised data collection requires us to look at the surveys, or more specifically what data requests should be in which survey, differently. It would be more efficient to adapt survey borderlines to data availability in order to be able to extract relevant information from business systems that might not always speak easily with each other.

Many enterprises are involved in a large number of surveys; this is especially true for enterprises within the manufacturing sector and other sectors where there are a limited amount of medium sized enterprises. Although there is no substantial duplication in the data collection, there is still information that is similar in different surveys, which is perceived by the enterprises as if we collect the same information several times.

At Statistics Sweden we have recently initiated a process that involves coordinating questionnaires by clustering variables to fewer data collections. This is something that will put great demand on how we save and store data as well as the methodology and design of surveys.

The basic idea of this coordination is to start from where the data requested is located within the company. An extreme situation would be one monthly survey, one quarterly survey and one annual survey. As there may be difficulties in linking data from different business systems in companies, a better idea could be to consider one survey for each business system or to start from what is available within a specific function in the company and collect all the requests linked to it in a minimal number of queries (monthly, quarterly and yearly).

Figure 3 describes the current situation where many surveys have their own data collection, but some survey information is also used to produce outputs in combination with other surveys. In figure 4, the survey borderlines are re-designed and data collection is organised more from the respondents' perspective and tailored to what information is available in different sources and business systems. For example, one survey could be collecting data from the accounting system through a machine-to-machine solution, whereas another survey could collect data and information that is not readily available in the system and needs manual treatment by respondents, therefore using a web questionnaire. Other data sources could be administrative registers from their authorities and even big data sources available in the "cloud".

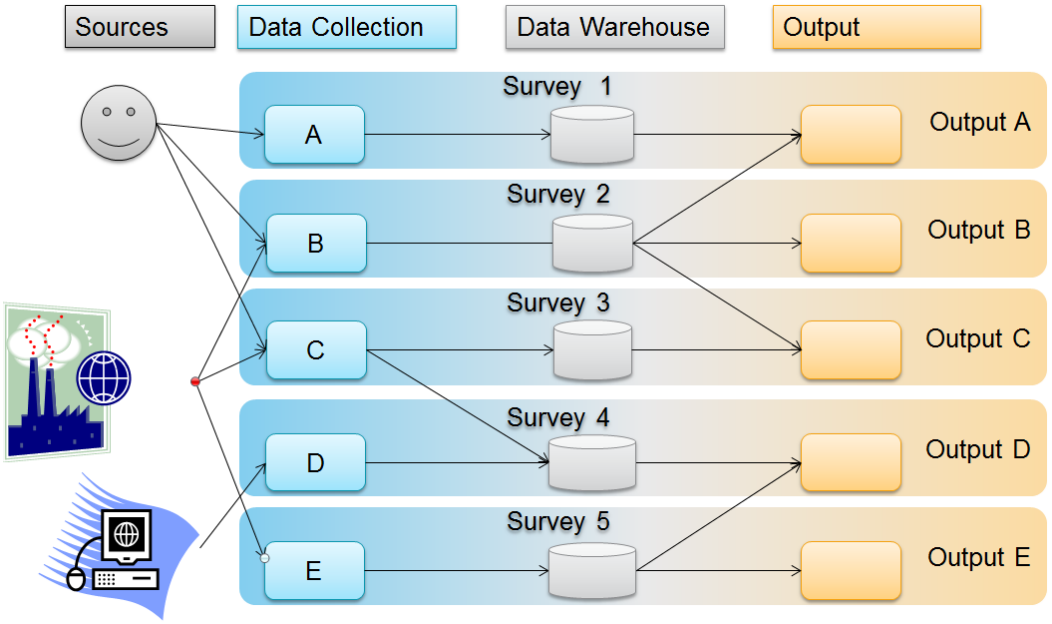

Figure 3: Coordinating questionnaires – The current situation

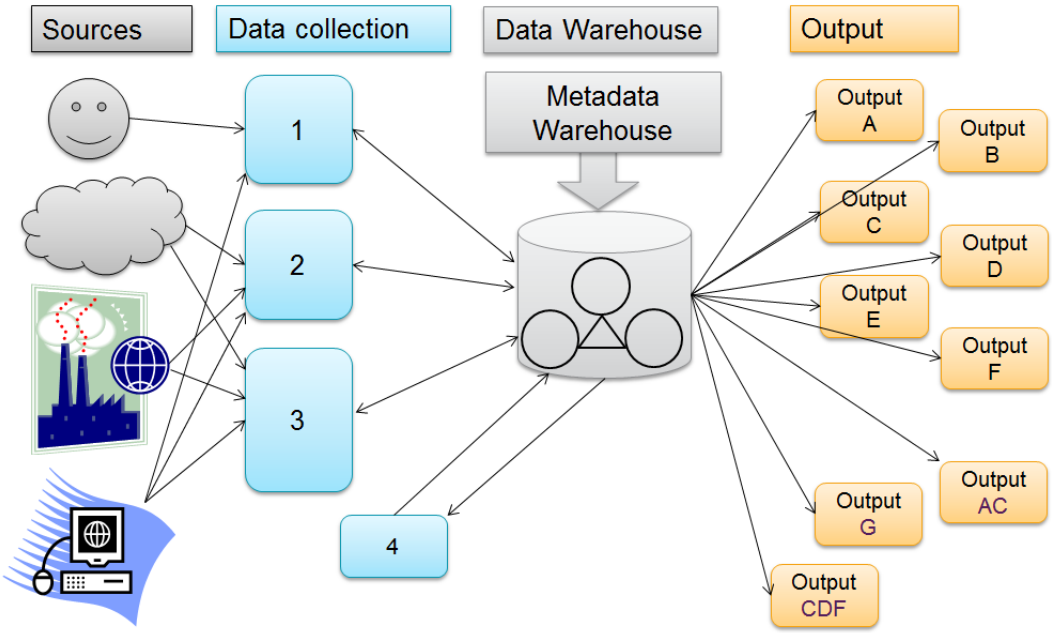

Figure 4: Coordinating questionnaires – The vision

3.3 Common technical solutions

To make use of the digitalisation possibilities, the Statistical office needs to offer standardised and robust technical solutions that allow for businesses or their IT service providers to set up machine-to-machine data provision or generate data for semi-automatic solutions. This involves formats, transfer methods, security and much more. For this to be interesting to businesses, the solutions cannot differ from survey to survey,

but must be common and easy to use. The statistical office also needs to offer support in this work, meaning that Statistics Sweden's IT department also needs to be prepared for this shift. It also means adapting the IT systems within the Statistical office to allow for large data flows coming this way, and new processing tools.

3.4 Co-operation with others

For digitalised data provision to be interesting to the data providers, both businesses and their service providers, it must be a worthwhile investment for them. And for it to become that, it is probable that they should be able to use the same data and processes for several provisions to different authorities. The Statistical office is in fact a rather small player in this game, for example relying much on samples and not total enumeration, which means only a smaller number of businesses are asked to provide data. Which enterprises are asked also changes over time, the largest businesses are more or less always included in the sample but not smaller businesses. So the business case for building digitalised solutions for statistics only is rather weak. But the same financial information is used to provide data for annual reports and for tax reports for example. So if the same information could be used for several data provisions, the business case becomes much stronger. If each authority that collects data sets up its own formats, its own technical solutions and uses its own definitions it will not be attractive to businesses and service providers like software producers to implement a digitalised data collection. But if the same information using the same definitions (the ones that are already in the business systems), the same format and the same technical solutions can be used to send data to many authorities, it will be much more attractive to build these processes into business systems. Therefore, it is very important that authorities work together on these issues and agree on concepts and definitions as well as technical solutions.

3.4.1 Coordination of concepts

From Statistics Sweden's experience, the coordination of concepts and definitions is probably the biggest challenge. Just like different surveys in the Statistical office have their history and are used to using their own terms and definitions, the same applies to different authorities. And if the authorities are very independent, cooperation with other authorities is not naturally built into the routines. But it is possible to come to agreements on concepts. To agree on content and definitions is a huge undertaking, but it still needs to be done. In 2015, a working group with representatives from the Swedish tax agency, the Swedish Companies Registration Office and Statistics Sweden made an investigation on how common concepts can be developed and maintained in Sweden (Brede et al 2015). The report was delivered to decision-makers within the different authorities in November 2015 with suggestions regarding future work. One way that has proven successful in Sweden and that was proposed in the report is to start with concepts defined by an authoritative agency, like the concepts in the accounting laws and standards, and to build common terms from these. Upon this foundation, authority specific concepts can be added as long as they are connected to the common concepts.

3.4.2 Formats

When it comes to technical formats, XBRL (eXtensible Business Reporting Language) is the format that is most likely going to be used for most business data provision. XBRL is an open international standard (based on XML) for digital business reporting, managed by a global non-profit consortium, XBRL International. XBRL provides a language in which reporting terms can be defined. Those terms can then be used to uniquely represent the contents of financial statements or other kinds of compliance, performance and

business reports. XBRL lets reporting information move between organisations accurately and digitally. XBRL is already used in more than 50 countries worldwide. In Sweden, the uptake of XBRL has been slow, but in 2015, an effort was made to change that. A technical committee lead by SIS, Swedish Standards Institute and consisting of the Swedish tax agency, the Swedish Companies Registration Office, Statistics Sweden and the Swedish Financial Supervisory Authority together with XBRL experts agreed on making XBRL Swedish Standard for electronic reporting of financial information. This means that the authorities have committed to introducing XBRL as specified format for providing data. This will hopefully have an impact on software companies realising that this is the format to build into business system to extract data for reporting to government agencies.

But XBRL is "only" a technical format; it also needs to be filled with content, definitions of reporting terms as described above. This content is called taxonomies. With a common understanding on the terms and definitions, a taxonomy structure with some common parts and some authority specific ones can also be developed. With such a taxonomy structure, it will also be possible to exchange information between authorities in an efficient way.

3.4.3 On-going cooperation

In July 2016 a government decision was taken to give the Swedish Companies Registration Office a task to prepare a solution for receiving annual reports in digital format (XBRL). The work will be carried out in formal cooperation between the authorities concerned (including Statistics Sweden and the Tax office). It is also stated that this work should prepare for the common concepts and taxonomies needed in this particular area (annual reports) to allow for an expansion of digitalised data provision to other financial information, and also to facilitate data sharing and reuse between authorities. This means the road to digitalised data provision to several authorities using common concepts and technical solutions has now formally been opened by the Swedish government, a starting point for a Standard Business Reporting program in Sweden.

4 Conclusion

Digitalisation is the next step in data provision from businesses to governmental agencies; Statistics Sweden is now in a transition period from web questionnaires to more automated forms of data provision. There will be a floating scale of automation, from semi-automated solutions to fully automated machine-to-machine solutions.

Statistics Sweden is working to support digitalisation and simplify data provision, but there is still some way to go before everything is in place. We need to think differently on how to organise our surveys and data collections and adapt to concepts and systems used by businesses for this to be efficient.

Besides internal work, there is also a great need to work together with other authorities to support digitalised data provision. This means a lot of work and challenges, but there are also large possible gains in efficiency and better data collection processes in the end. The era of authorities working by themselves has come to an end, cooperation and collaboration will be the only way to move forward. And the government has now opened that road.

References

Bolagsverket, Jordbruksverket, SCB, Skatteverket, Transportstyrelsen: Lätt att lämna rätt - En myndighetsgemensam rapport om hur det kan bli enklare för företag att lämna ekonomiska uppgifter till myndigheter. (2013)

Brede, N, Persson, A, Erikson A-G, Levinsson, L, Rastman, M, Reinsson, C, Schröder, M: Standardisering av finansiell information – en möjlighet för Sverige (Bolagsverket, SCB, Skatteverket, 2015)

Adolfsson, C, Hertzman, C: Developments in collection modes and contact strategies in business statistics (2015)