**Data sharing in the public sector**

*– A pilot for the Danish universities*

Annie Stahel, Statistics Denmark, ast@dst.dk

**Abstract**

*In Denmark, the legal basis for collecting data for official statistics is ensured by the law on Statistics Denmark. Data that has been collected for the purpose of official statistics may not be used for other purposes, e.g. administrative purposes on a micro level where it is possible to identify the individual. The initial owner of administrative data (i.e. other public authorities) can, however, authorize and instruct Statistics Denmark in the role of data processor in different ways.*

*This paper investigates data sharing from a legal, organizational, and technical viewpoint and shares an example of concrete data sharing with the universities in Denmark. The paper will clarify what Statistics Denmark considers data sharing to be and describe the considerations that have been taken in connection with the sharing of data. We realize that the topic of data sharing has the potential to cause a certain amount of concern in NSI’s as it did at Statistics Denmark and this aspect is also considered. The end result is a new way of utilizing competences concerning statistical micro data processing methods for non-statistical purposes in an increasingly data driven public sector.*

**Keywords:** Facing the future, new roles and opportunities for the NSI’s, data sharing, data processing, administrative purpose

**1. Strategy and core values at Statistics Denmark**

For some time, Statistics Denmark has had thoughts about how to facilitate a new and useful use of data and the data processing capabilities that we possess. With data being one of the world's most valuable resources, it is only fair that Statistics Denmark contributes to society in this field. Therefore, data sharing is an integral part of our Strategy 2022.

The core values of Statistics Denmark are independence, trustworthiness, data protection and user orientation. But as an old institution with its roots stemming from the birth of the Danish democracy we also need to develop in order to remain relevant in a continuously evolving world. Therefore, we want to strengthen our competencies and activities, especially in areas that can be characterised by our 3 new values, namely adaptability, a holistic approach and openness.

It is clear that some of the values can appear to clash at times. For instance, how do you develop openness with regard to e.g. data when your first priority is keeping data secure?

Therefore, a fair bit of work and considerations have been put into the topic of how to share data (in a very specific sense) without compromising data security and violating the trust of the users. We believe that we have found a model.

**2. Data sharing – a rather delicate subject**

We realize that there are probably several perceptions of what data sharing is. In this context, we define data sharing like this: data sharing is a process where one data collection channel is used for collection of data that afterwards can be utilized for either statistical and/or administrative purposes in two different and segregated environments. The data sharing must be carried out in an independent and secluded area in order to prevent mixing up of the roles and responsibilities as data processor and data controller, respectively.

The fundamental principles and underlying premises are that data that has been collected for the purpose of statistics may not be used for other purposes, e.g. administrative purposes. Data collected for administrative purposes may, however, be used in other ways too, including statistics.

The notion of data sharing is in itself a delicate question that often causes a bit of concern amongst employees who work with the production of statistics and who have data protection as an integral part of their DNA. Is it vital to note that data from the statistical production area is not shared with anyone outside the NSI. The key aspect of data sharing in this sense is the pre-processing and exposition of the received data *before* it is transferred to the area of statistical processing.

Two major factors have triggered the entry of Statistics Denmark onto the data sharing scene:

1. there is a substantial pressure from official parts of Denmark on Statistics Denmark to ‘open up’ and make use of the substantial assets that we possess, both with regard to the skill set as well as the unique sets of data
2. Statistics Denmark has ambitions with regard to the value aspects and societal contribution and actually wants to have an impact on the debate and value creation in society

**3. The universities and their need to inspect their own data**

For some time, the management in the division of Population and Education has had talks and discussions with the Ministry of Higher Education and Science about the data in the educational field, covering the universities as well as other kinds of institutions of higher education. The situation is that the universities deliver data for statistical purposes once a year to Statistics Denmark on the ordinary legal basis, so to speak, where after Statistics Denmark performs the statistical editing and needed correction of errors. The resulting data set with data on a micro-level is then handed over to Ministry of Higher Education and Science, which in turn will provide the universities with calculations that establish the extent of their funding and financial contribution. The universities do not see this data on a micro-level and may never do so because they are not allowed to make decisions about the students on a micro-level based on the information from Statistics Denmark.

One of the problems with the current solution is that the universities themselves are not always aware about errors and ‘ghost’ students who should not have been enrolled because they are associated with another university. Statistics Denmark on the other hand has the optimum conditions for sorting out the population of students and determine the proper allocation of them based on the totality of data from all the universities (and other educational institutions).

This insight and corresponding editing process in Statistics Denmark has resulted in the so-called gap-closing student register. This register is probably unique in the world since it has a full and un-broken record of every person from the moment they enter the educational system as preschoolers to the moment they leave university (…If they take this educational course, that is).

In order to be able to understand their individual allocation of financial resources, the universities would benefit greatly if they knew what the basis of the calculation was. This means that they would actually like to get insight into the edited and error corrected set of data that Statistics Denmark has produced about their students, on a micro-level for each specific student.

Therefore, an agreement was made between the Ministry of Higher Education and Science and Statistics Denmark in 2018 that we should develop a data sharing solution where the universities could gain insight into their own data without compromising the data security across the university sector and without violating the regulations that Statistics Denmark works under.

The solution is to have Statistics Denmark work both in the role as data processor as well as data controller in different steps. After the first step and after the initial editing Statistics Denmark will deliver data back to the Ministry of Higher Education and Science, which in turn will allow Statistics Denmark to share this data with the universities by means of a data processing agreement. The universities will get access to the identifiable micro data about their own students, their results and study program.

The second step is to forward the set of data into the ordinary statistical process which will result in the statistical products as they have been known and used for decades.

**4. A model for data sharing: segregation of environments, roles and responsibility**

Statistics Denmark decided early on to engage with a law firm in order to make sure that all possible considerations were taken and all risks investigated.

The law firm in case, Bech-Bruun, provided legal advice concerning different organisational and technical aspects of data sharing with the universities during the end of 2017 and throughout 2018.

With regard to the organisational aspects several set ups were considered; please see figure 1.

**Figure 1: Different solutions for environment for data sharing with universities**



The first and perhaps most obvious solution (no. 1), namely to establish a data sharing environment within the area for statistical production is not a viable option. The reason for this is that the status of Statistics Denmark as a data controller in its statistical editing role should not be mixed up with the role as data processor for an external authority.

With regard to the other two options, which are:

* no. 2: to establish a totally new “data factory” with its own environment, separated by a firewall and equipped with an immaculate set of new tools and licenses or
* No. 3 to establish a data editing environment in an existing area, segregated by a firewall segment of its own; the environment at hand is the research area for researchers.

For this pilot, no. 3 was chosen because establishing a brand new environment from scratch as in no. 2 would have been a very costly activity – even if it would be carried out with the funding from the Ministry of Higher Education and Science.

**5. A two-phased model for data sharing**

The model that all involved parties have agreed upon, meaning both the Ministry of Higher Education and Science, the law firm and the relevant committees and management at Statistics Denmark can be illustrated as in figure 2:

**Figure 2: Set up with Statistics Denmark in a data processing and data controlling function, respectively**



This model ensures a strict segregation between the delivery of data for statistical purposes and for data processing and sharing, respectively. This is enabled via the separate IT environments with a clear separation between data for statistical purposes and the area for data for processing with data sharing as the end goal. This way, it is possible to utilize the expertise that Statistics Denmark possesses with regard to data processing, validation, correction etc., and thereby deliver a valuable service back to the society, in the present case the universities.

The legal staff at the Ministry of Higher Education and Science have approved this construction. The universities have agreed upon it, and they are very eager to get access to their edited and corrected data.

**6. Data sharing within the organisational boundaries of Statistics Denmark?**

The law firm advised Statistics Denmark to investigate whether it would be possible for Statistics Denmark to establish a separate legal entity that would be in charge of the data sharing and data processing; a new sister company, in a way. The purpose of this would be to make it very clear that the segregation between the statistical production environment and the data sharing environment is absolutely rigorous and trustworthy.

It is, however, not in the power of Statistics Denmark to engage in the creation of other legal entities. All activities related to the work of Statistics Denmark must be carried out within the framework of this institution. Among such activities, data processing for other parties is actually specifically mentioned in the law text concerning Statistics Denmark.

**7. The technical solution**

The environment for data sharing is based on a server set up in the secluded area for researchers; this area is governed by a separate Active Director with separate users and log on profiles, and no confusion with the statistical production environment is possible.

Data is channelled into Statistics Denmark through the Central Receiving Platform and after that allocated to an Oracle database. The data is then moved directly to the data sharing environment before any specific data processing has taken place.

After an editing process which is not geared for statistical purposes but for error correction, the universities will get access to a file with their own reported data.

The data editing process will, for instance, include the following operations:

* All dates are tested for correct and logical form and content.
* All individual students who have two registrations at two different universities will be edited; time overlap will be solved and one record possibly deleted - we take the latest starting point in time as the most reliable.

The student variables with the edited and corrected data together with a clear marking that indicates whether the record has been edited or possibly also deleted in the data processing process is then presented for the universities together with their own initial file.

**Table 1: Example of a summary file for a Danish University**

|  |  |
| --- | --- |
| stat\_txt | number |
| INSTITUTION: University of Copenhagen  |  |
|  No. of records: | 51,295 |
|  Hereof no. of edited records: | 9,848 |
|  Hereof no. of deleted records: | 194 |
| Variables: |  |
|  No. edited PBGTP: (START) | 612 |
|  No. edited AFGTP: (END) | 421 |
|  No. edited UDD:(UDD refers to the study program) | 8,565 |
|  No. edited AUDD: (COMPLETION) | 0 |
|  No. edited UDEL: (EDUC.PART) | 566 |
|  No. edited INSTNR: (INSTITUTION) | 0 |

Data files will be returned via a secure FTP platform to individual users of the different universities, and a summary file is presented cf. table 1.

At this very moment, the Ministry of Higher Education and Science is evaluating the solution and the results, including the summary files as shown above in table 1, before the solution becomes accessible for the eight universities.

After that, the solution will go live for the universities and the idea is to evaluate the process and the degree of success before the next step will be considered: opening up for data sharing for the rest of the educational institutions involved in higher education in Denmark.

The price for this is just under 100,000 euros in total with about 32,000 euros pr. year in a 3-year period including the installation and establishment of the technical environment.

**8. Perspectives and future work**

Our hope is that the rest of the educational institutions will be able to have their share of the data either by the end of 2019 or by the beginning of 2020.

And what comes after that? The ambition is to make a pilot project and see if we can build a specific platform for data sharing based on open source based software. This way, the costs will be reduced significantly in comparison with the existing platform that we have with software solutions from commercial vendors. We are planning on using PostgreSQL as a database tool and R and Python for programming and data handling. The end goal is to try to make the third option for data sharing come into existence as shown in figure 2 – but without the license expense.

**Figure 2: Generic environment for data sharing**



**9. References**

References below refer to mostly internal papers at Statistics Denmark.

Osborg, Liv Palmelund et al. (2018) Illustration af Danmarks Statistiks om databehandler (English: Illustration of Statistics Denmark as data handler) [BECHBRUUN: case no. 059991-0004 mif/lpos/chko Doc.no. 19595526.1]

Osborg, Liv Palmelund (2018) Vurdering af forslag til datadelingsløsning [BECHBRUUN-Legal.FID3460479] (English: Assesment of proposed solution for data sharing)

Stahel, Annie et al. (2018) Indstilling om datadeling – sådan sikrer vi data (English: Proposed solution for data sharing - how we secure data)

Sørensen, Susanne Mainz, (2018) Dannelse af det hullukkede elevregister, som stilles til rådighed for universiteterne (English: Formation of the gap-closing student register made available for the universities).