United Nations ECE/CES/2023/21

Economic and Social Council

Distr.: General
2 June 2023

English only

Economic Commission for Europe

Conference of European Statisticians
Seventy-first plenary session
Geneva, 22–23 June 2023
Item 3 of the provisional agenda
Moving towards open-source technologies – strategic and managerial perspective

"DAPLA" – a cloud-based statistical production system and its implications for Statistics Norway

Prepared by Norway

Summary

National statistical offices (NSOs) are facing increased competition, higher user expectations, fast paced technological developments and a changing data landscape. To stay relevant and fulfil their mission there is a need to keep up with the demand for more and faster statistics, including being able to quickly respond when crises occur. At the same time, funding is limited and NSOs needs to carefully prioritize and optimize their production processes to save costs and/or increase their capacity to meet the needs of their users.

One possible move is to utilize open-source technologies. It is not clear to what extent this can help NSOs resolve challenges ahead. Moving from one programming language to another, such as replacing SAS with Python or moving from on-premises IT-systems to the cloud, might seem like minor changes of little consequence for the fundamental strategic challenges.

This document discusses the experience at Statistics Norway (SSB) where they do find the implications of such changes to be both far reaching and with a possibly profound impact on the business, but also hard to manage. Some of the trends that create pressure for change and provide new opportunities are described before looking at where Statistics Norway is heading. Finally, the challenges and dilemmas that Statistics Norway has experienced and how these are handled are elaborated.

The document is presented to the Conference of European Statisticians' session on "Moving towards open-source technologies – strategic and managerial perspective" for discussion.

I. Introduction

1. The current Norwegian Statistics Act[1] was adopted by the Parliament in 2019 and entered fully into force on 1 January 2021. According to the Act, Statistics Norway is the central authority responsible for coordinating the development, production, and dissemination of official statistics and for reporting on the quality of official statistics.

2. An important feature of the Norwegian statistical system is that official statistical production is to a large degree based on administrative registers and the mandate to use register data has been strong for decades. Statistics Norway has access to more than 100 administrative registers from almost 30 public institutions which are utilized to produce official statistics.

3. The new Statistics Act also specifies access to other data, such as privately held data, for the development, production, and dissemination of official statistics. Utilizing new data sources is one of the strategic goals for Statistics Norway in order to provide better quality statistics and to remain relevant, an objective also emphasized by both the Ministry of Finance and the Council for Statistics Norway[2]. Increased access to data present opportunities by providing data on new areas and phenomena as well as these may be more granular as well as more frequently updated.

4. At the same time, there are also challenges when utilizing these data for official statistics. Traditional public sector registers change in both structure and data over time, and such changes may occur even more frequently in the future as register owners are improving their services for the users. Such changes may be even more frequent for privately held data. Surveys continue to be an important data source, but declining response rates and skewed responses cause challenges. Data sharing is increasingly important with raising expectations[1] that Statistics Norway offers new methods of sharing data while preserving confidentiality. Finally, the General Data Protection Regulation (GDPR) and cybersecurity remains areas of continued focus and investment.

5. Many actors recognize the value in presenting their version of societal development, and dissemination is faster and spread over more channels than ever before. Some may also present numbers that resemble official statistics. The diversity of information means more people have access to knowledge, but it can be difficult to know the quality of the analyses being disseminated and to distinguish information from disinformation. Technology and digital products develop fast, and user expectations are ever increasing.

6. To handle such challenges, NSOs need to make it easier to benefit from general technological progress, also when it comes to collecting, transforming and re-using data.

7. At present, Statistics Norway holds its own data centre which hosts all hardware and software required for the current production processes. The system portfolio consists of many different solutions, most of which has been developed by Statistics Norway over the years. The underlying technology (Linux, Oracle, SAS[3], FAME[4] and others) is robust, but some systems are complex and resource-intensive to change, and thus managing the total portfolio of systems and integrations is challenging. SAS is widely used to transform data in the statistical production along with the above-mentioned portfolio of IT-systems that are mostly operated through graphical user interfaces (GUI).

II. Where are we heading

8. Statistics Norway is currently in the process of building its own Data Platform ("DAPLA") in the public cloud (Google). The data platform provides common services such as security, storage, data sharing, pseudonymization, and to produce statistics with the

[1] https://www.ssb.no/en/omssb/ssbs-virksomhet/styringsdokumenter/statistikkloven

[2] https://www.ssb.no/en/omssb/ssbs-virksomhet/organisasjon/radet-for-statistisk-sentralbyra

[3] SAS AF and SAS EG. SSB has decided not to implement SAS Viya and are aiming to reduce the current use of SAS to save costs.

[4] https://en.wikipedia.org/wiki/FAME_(database)

programming languages Python and R. Over the next few years, the aim is to move all data to the public cloud and to re-create all statistical production processes from scratch. The python or R-code required is developed by the statistics departments.

9. Some factors regarding DAPLA that could contribute to increase the development speed and ability to respond faster to changes in demand for new statistical products are that:

- DAPLA provides the statistics production teams with more opportunities to enhance and control the flow of data end-to-end than the current on-premises systems.
- DAPLA provides easy sharing and re-use of data to support analysis and development of statistics.
- DAPLA provides access to some of the leading open-source technologies:
    - Open-source languages such as Python are extended at a high pace and offer libraries for virtually any conceivable problem[5].
    - Jupyter Notebooks offer the integration of documentation, mathematic formulas, illustrations, and executable code. Notebooks also offer collaboration mechanisms allowing employees to work together on the same code base in real time.
    - Github has emerged not only as an excellent version control and storage facility for code, but also provide strong support for development processes such as agreeing on new features, how they should be implemented, who should do the implementation and who will approve. Github today is not only the world's largest code repository but is also an innovation hub enabling code sharing and re-use and fostering innovation networks where organisations collaborate across industries and geographical borders.

10. It is difficult to calculate the potential for optimizing the current production processes and therefore this remains an area open for discussion.

III. Challenges and how they are handled

11. There are quite a few dilemmas and challenges associated with the approach outlined above; best way to develop new competence; solve the cost, time, and resource constraints; introduce an agile mindset to development, and it remains to be seen to what extent Statistics Norway will succeed in achieving the goals. Some of these are discussed in this section.

A. Competence

12. As change has become the normal state of play, the idea that the statistics production teams must own their production process end-to-end and have the tools, skills and authority to make changes as required may seem evident. However, it remains to be seen how successful Statistics Norway will be in scaling this approach to cover all statistical production as we are still at an early stage.

13. Transforming and manipulating/editing data through code has been part of Statistics Norway's statistical production for decades. The SAS programming language has been the most important, but also SQL, Stata, R and other languages have been used. FAME is an important tool both for analysing time series in short-term statistics as well as to produce some macro statistics (e.g., financial accounts statistics). The importance of coding skills and the extent to which it is expected to be part of day-to-day work will, however, increase in the coming years.

[5] One interesting example is the recent effort by SSB economist Magnus Helliesen who has created a python program that resolves an arbitrary collection of equations. In only 800 lines of code it resolves the 15500+ equations constituting the Norwegian National Accounts in seconds. https://github.com/statisticsnorway/model-solver#readme

14. Some important initiatives have been taken to support the transition, and the HR division is playing a crucial role in this.

(a) Competence mapping and follow-up: A structured approach has been developed to map employee skills against the skills Statistics Norway is expected to need for the ongoing transition. This is different from the traditional competence stock taking where emphasis has been on what competencies employees have, rather on identifying what needs to be developed moving forward. Work is also done to improve skills-based career paths.

(b) Agency School: Statistics Norway's internal course programme is playing a crucial role in enabling the transformation. It is offering a mixture of MOOC[6] courses and internal training. Regular events such as "buns and coding", "Friday Digitalization", and more are organised. A recent event owned by the statistics departments and facilitated by HR division was the "DAPLA day", a full day with several short seminars providing information and discussions on DAPLA, open to all employees. A DAPLA-curriculum for employees has been developed and a curriculum for leaders is under discussion.

(c) Leadership-network: Middle management meet in smaller discussion groups facilitated by HR to share experience and discuss challenges.

In addition to building DAPLA itself, the IT department is supporting the transition by offering tutorials and documentation, providing technical and agile coaches, and leading an advisory team on good coding practices with participants from the statistical departments.

(d) Digital professional communities: Alongside the introduction of DAPLA, Python, R, notebooks and the like, the use of collaborative tools such as Yammer and Slack have multiplied. There is a lively exchange of questions, answers and discussions related to the various topics. A few early movers and some leadership attention is important to spark initial activity.

15. As the above brief overview suggests, much investment is done in human capital and mobilizing the organisation for the transition. One important aspect of this is the importance of leaders at all levels of the organisation taking the responsibility to lead, explain the way forward, why we need to change, how we execute the change, and what it means to the individual employee. This conversation should take place with individual employees, in groups and in larger assemblies.

B. Time and resource constraints

16. Until a statistic has been re-created on DAPLA it still needs to be produced in our current systems. This means that the statistics production teams need to develop on DAPLA and produce on-premises at the same time and invest much time in competence building in coding and other related "DAPLA-competencies".

17. Although the initiatives on competence and organisation outlined in the previous section are important, the most critical initiatives are those taken by individual middle managers to allocate time so that employees can learn-by-doing as part of normal day-to-day operations. This is hard and requires motivating employees and working together to find practical solutions. As there is no certain knowledge on what the most efficient method is, and local adjustments must be done to consider the production plans of the individual statistics, several approaches are being tested, and experience is shared between the statistics departments.

18. Some of the approaches being used in the statistics departments are listed below:

- Create some initial successes and examples that can inspire others
- Support early movers

[6] MOOC – massive open online courses. Universities and others offer high quality and low-cost courses online, see e.g. https://www.edx.org/search?q=python for courses from MIT, Berkely, Google and others.

- Utilize less busy periods for competence building and coding
- Create expert or innovation groups that work across the statistics production teams and
- "Carve out" pockets of functionality/small parts of the production process that can be re-coded in R or Python and preferably in such a way that a small improvement is gained either in quality or need for manual effort, thereby freeing up time.

19. To lower the barrier to the public cloud and DAPLA, a "mini-DAPLA" has been established on-premises. This makes it easy to access production data through notebooks, R and Python. Thus, current production pipelines may be improved whilst at the same time they become easier to subsequently move to DAPLA.

C. Managing costs

20. Current IT expenditures in our own data centres are mostly fixed and will to a large extent continue their presence until the last statistic has been moved to DAPLA. At the same time, cloud costs are increasing as we use more and more cloud services. Another cost related issue is that salaries for in-demand skills such as code-savvy economists and IT professionals with cloud expertise are increasing.

21. Statistics Norway is taking several actions to handle the cost problem including holding back on operational expenditures to be able to fund development activities, reducing the use of external IT consultants replaced by recruiting IT developers, and reducing IT operations expenditures on-premises by switching from proprietary software to open source for certain tasks such as monitoring. Measures are also taken to automate more of the on-premises IT operations processes and reduce the number of systems used in current statistical production, including scaling down the use of systems such as SAS and FAME. More extensive measures such as launching a general cost-cutting initiative has not been discussed.

D. Agile – the perception of change

22. Quality framework such as the Generic Statistical Business Process Model (GSBPM) within statistics and Information Technology Infrastructure Library (ITIL) within IT, are built on the underlying assumption that quality is achieved through well-defined processes where humans repeat the same steps in each production cycle or event. In addition, traditional waterfall IT-projects, which build on the idea that development is carried out within a given scope and timeframe, are still carried out. Under this paradigm, change is an exception and stability the normal and desired state.

23. With agile mindset and code as the means of production, this changes; quality is achieved through the continuous improvement of the code and change is the normal state. Much of the research and efforts in the agile movement goes into making the development-to-production cycles faster and more frequent.

24. While the idea of autonomous teams doing continuous improvement of their products and production process, without the need for detailed up-front specifications of what to make, is attractive for some, there are also plenty of pitfalls and risks when moving to agile at scale.

25. Statistics Norway launched a few pilot teams to get experience with agile. Initial results were mixed, and we learned that care must be taken to ensure that agile approaches do not develop into more bureaucracy than before. There has also been much confusion and discussion around new roles and concepts such as "product owner" and "autonomy" and how this interacts with the leaders' responsibilities and formal chain of command.

26. One important message from the director general and the director's group has been that we are working to identify and develop Statistics Norway's model for development. Whilst reading up on books and research, and learning about agile is important, establishing a development model is an explorative effort based on trying out new ways of working, learning, and adjusting. However, it is important that clear directions are given from the senior management; agile mindset, operations, and development should be melting together,

development should be anchored in the statistical production units, and all this should be done in a way that is practical considering ongoing production processes.

27. The concept of team autonomy may be easily misunderstood. In traditional development processes, much effort is put into specifications. Expected *output* is defined up-front. Also, hours worked and detailed project plans tend to be the order of the day. Agile on the other hand emphasise focus on *outcome*, meaning how a development initiative is expected to contribute to realising the company strategy, and how this can be measured. When team outcomes and successes can be measured, the team can utilize its competence and what it learns from users to decide what do to in the next iteration ("autonomy"). In short, a team need to have it clearly defined what it should accomplish, including both effects for users and for Statistics Norway, and also how success is measured, while leaving it up to the team how to realize these effects.

28. To assist in the process of utilizing agile, Statistics Norway has established a centre of excellence ("IT-partner") providing guidance and advice on agile processes.

E. Common functionality

29. Today's portfolio of IT systems offer functionality for statistical production that are used by many statistical products. Some examples are survey-systems, micro and macro editing systems, population registers and business registers. Required common functionality must be established in the cloud. This offers the opportunity to re-think current processes whilst there is also a risk that organisational inertia create pressure for just re-creating existing systems and processes with newer technology.

30. In general, deciding on what should be common and what should be local is considered to be one of the harder problems within the realm of IT. Code that is made solely for the need of a particular issue tends to be simpler than code made to cater for the needs of the many. This means that the code can be written and tested faster, is easier for others to read and understand and is easier to maintain. On the other hand, code that can solve a problem for many can also increase speed, reduce the possibility of error, and reduce costs. Whilst IT architectural principles such as "high cohesion and loose coupling" can assist, it is even more important that the organisation understands how common functionality and re-use is achieved in a code-heavy environment. Boilerplate[7] code, functions, libraries, and micro-services are some important examples.

31. Self-service for statistics producers is a very important aspect to consider when adding new functionality to DAPLA. As self-service on-premises and on DAPLA is quite different, there is a need to increase "DAPLA-competence" among the statistical production staff and "statistics competence[8]" among the IT staff. It is important to facilitate meaningful conversations on what should be the common functionality, how this common functionality should be materialized (e.g., library and microservice) and what should be left as the responsibility of the individual statistics production team to avoid copying the centralized, hard-to-change, monolithic systems that underpin current production processes.

32. Currently, all statistics departments have plans for re-creating their statistical production processes on DAPLA. Plans are kept at an overall level, and timeframes are typically divided into near, intermediate, and long term, with less difficult statistical productions scheduled to start first.

33. As DAPLA is also continuously developed at the same time as the new statistical production pipelines are developed, a high-level DAPLA-roadmap is maintained for predictability and as a basis for discussions about priority. The DAPLA product owner maintain dialog with the DAPLA-users to uncover needed functionality.

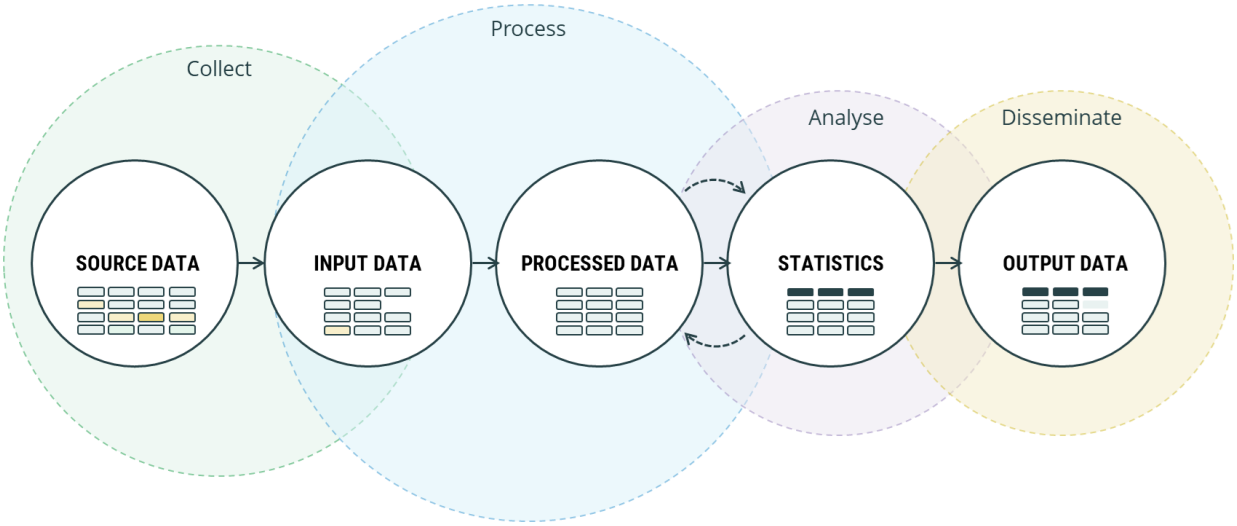
[7] Standardized sections of code that serves as a template for creating new code. It helps developers save time by reusing established patterns and practices.

[8] For example, understanding the processes associated with maintaining the business registry or the national accounts production process.

34. A *Development Committee* has been formed to prioritize scarce IT resources across Statistics Norway. The committee has representatives from all the three statistics departments, from the methodology, communications, and research departments and is chaired by the Director for IT. One important initiative is a structured investigation into possible common functionality. Special working groups are also in place for in-depth work on special areas such as the business registry or how editing can be solved.

35. The (*Statistical) Standards Committee* at Statistics Norway has been revitalized. Whilst new and updated standards apply to both on-premises and DAPLA production pipelines, it is the transition to DAPLA that really provide the opportunity to implement the standards. The most important standard is the data steady states that can be seen as an inverted GSBPM where focus is on the resulting data products rather than the processes, supporting the concept that improvements in quality and production is achieved through ever changing and improved code.

Figure 1
Data steady states versus GSBPM



Source: Data Steady States standard, Statistics Norway

36. DAPLA is built around steady states and teams, providing each statistics team with a structure to provide for production pipelines that are easily recognizable across all statistics. All teams have one or more data administrators and only the data administrator may access the source data as they are seen as the most sensitive state. The steady states provide a fundament for teams to implement quality and production control. Also, meta-data and data sharing mechanisms are closely tied to the steady states.

37. As more work is invested in building statistics on DAPLA, pressure for common functionality whether in code, services, or methods, is increasing. DAPLA must offer services needed so that individual statistics production teams do not have to re-invent the wheel and build too much themselves, and at the same time teams must be able to control and improve their own production process without too many dependencies on central IT-solutions and IT-developers. Overall, Statistics Norway must avoid unnecessary complexity and costs and ensure that possible gains in efficiency and quality are materialized. Striking the balance between what is centralized and what is left with the individual statistics production teams is crucial.

IV. Conclusions

38. Moving towards open-source at Statistics Norway means moving all data, all production processes and the organization of this endeavor. This is in some respects a daunting task. But it is also something that Statistics Norway and other NSOs have done in the past. Programming languages and computers have changed much over the years, but the fundamentals of coding have remained stable. A fluent SAS-programmer is well equipped to

master new languages, and continuous improvement and development has always been on the agenda in statistics departments. Statistics Norway has taken a holistic approach, and as we are learning and mastering new skills, the employees at Statistics Norway continue to be our most valuable resource. Our best investment towards success is by creating the right conditions for employees to thrive and continue their development.