Privacy by Design in Official Statistics: Should it Become a Default/Standard?

Baldur KUBO, Angela SAHK, Veiko BERENDSEN
Cybernetica ; *Statistics Estonia*

Privacy by design (PbD) as an approach to systems engineering has been conceptualized for more than a decade. It has inspired the legal norm of managing personal data in EU and incorporated into the GDPR. In practice PbD is far from being the standard in systems engineering in both private and public sector procurement.
We explain why privacy needs to be engineered into statistical systems with sample risks and hazards when privacy is patched into a solution. We showcase the PbD approach of future proofing existing services with a redesign example of an official tourism statistics application that uses mobile location big data.
Finally we discuss how the statistical system could proceed in adopting PbD and build trust required for capturing new data sources.

1. **Introduction**

Capturing, storing, transforming and analysing confidential or personal data is part of the daily routine of every national or transnational organization producing official statistics. We are presently living in the time of exponential growth of digital data being generated by enterprises, people and devices, where new data and data sources are becoming candidates for simplifying production of official statistics. The granularity, volume and impact of such data is growing in orders of magnitudes as we move from annual surveys and corporate reports to real time data, capturing states and transactions between every organization and person: finances, locations, social interactions, health and emotional state of everybody all the time.
According to the IDC Data Age 2025 report 90% of the data generated within the global datasphere will need securing, but only half of it will be secured by 2025 [ref 1]. In order to not find oneself in such a situation where half of the data under work is not secured, security and privacy need to get attention of executives of statistics organizations.
As statistics organizations are starting processing big data sources where the granularity and criticality of data is much higher than what has been available from surveys or annual reports until now, privacy becomes a high priority. For this evolution we need to have safeguards embedded into our data management processes to protect fundamental rights of citizens and organizations. These safeguards may include fundamental changes to how statistics organization produces

statistics. For instance outsourcing part of the aggregation process to data owners, cooperating with data owners using distributed secure computing or securing individual data records with cryptographic means during the whole data processing lifecycle. That is where privacy engineering enters the picture, to help select the optimal management and technical controls.

2. What is privacy engineering

NIST defines *privacy engineering* as the "discipline of systems engineering focused on achieving freedom from conditions that can create problems for individuals with unacceptable consequences that arise from the system as it processes PII." [ref 2].

PbD is an approach to systems engineering that recommends privacy to be considered throughout the whole engineering process, starting from initial design. It helps us in moving from high level policy requirements like "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes" [ref 3] or "Access for statistical purposes to other data, such as privately held data, is facilitated, while ensuring statistical confidentiality and data protection." [ref 4] to the choice of management and technical alternatives in design. For software engineers PbD is part of the risk reduction within the software development process.

For the legal system, the current European General Data Protection Regulation has been inspired by PbD. The second principle of PbD 'data protection by default' and 'data protection by design', is incorporated in GDPR and thus the current legal norm. Though the producers of official statistics may be exempt from some of the requirements within the GDPR, the owners of innovative new data sources are restricted by GDPR in sharing or repurposing their data for statistical purposes.

Executives are able to make managerial choices when they have a clear picture of the alternatives and impacts of their choices. Let us illustrate the criticality of involved managerial decisions by risks if privacy is not engineered into the system [ref 2]:

- Re-identification of information linked to an individual;
- Misunderstanding the extent to which they have consented to share their data;
- The perception of a loss of privacy changing behaviour;
- Embarrassment when information tied to the individual is released;
- Information is used to discriminate against a group of individuals;
- Information is accessed by law enforcement.

Risks concerning organizations if privacy is not designed into the system may include:

- Financial losses to the concerned organization;
- Damage to reputation to the organization and producer of statistics;
- Loss of market share;
- Change of customer behaviour, including churn and reduced cooperation with the statistics organization;
- Loss of market value, damaging all shareholders;
- Potential litigation of the producer of statistics.

The majority of these risks can be mitigated and the involved data be secured when PbD is considered early on, during the whole lifecycle of the statistical production system.

3. Why cannot privacy be patched into systems?

Privacy is a fundamental requirement for a system implementing a statistical workflow and therefore "patching it in" is challenging. Attempting to add privacy later might require parts of the system or the entire system to be re-designed and rebuilt to guarantee privacy. Key threat situations might be missed if the system documentation is not up to date or does not include all information (often it does not). Avoiding the necessary redesign might lead to the use of unsuitable privacy technologies.

As a simple everyday example, let us consider the situation of buying a used apartment where renovation is required. Whilst carrying out the renovation, the future inhabitant is still living in a different location. After they move in to the newly renovated apartment, they discover that the walls must be very thin, because in the evenings they can hear the neighbours talking to each other, even without raising their voices. The issue was not identified prior to the renovation as during the day the neighbours were not at home as they were out working. But if you can hear them, they would most probably hear you as well. This disturbs the inhabitants heavily and you need to find a solution.

Options seem to be to sell the apartment and find a new one; sacrificing privacy and doing nothing or re-renovating the apartment again, making the outer walls soundproof. None of these options are good and include either sacrificing privacy or finding additional resources to solve the privacy issue. Either a lot of time and effort needs to be put into finding a new apartment or more work is needed than just add another layer of sound-proofing materials on to the outer walls, i.e parts of the ceiling or floors may need redoing. The most economical solution would have been considering privacy requirements whilst finding the apartment and before renovating it. From the moment of

finding out the privacy problem and solving it, we have an active privacy issue, bringing inconvenience and potential waste of resources.

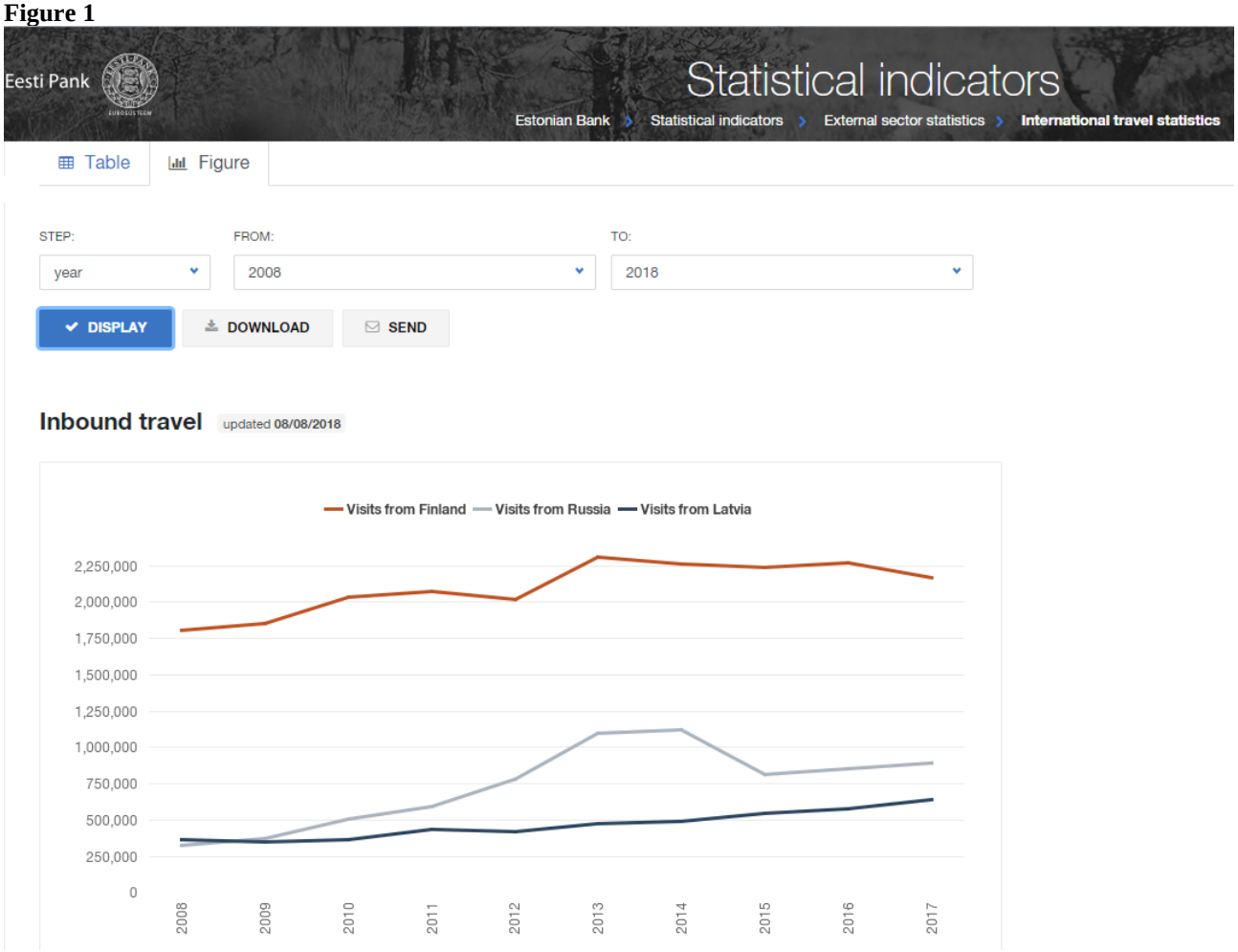
In the design of data analysis workflows the similar situation is not uncommon. It is easier to start working with open data compared to handling privacy and confidentiality requirements of confidential data sources. Thus the experimentation and tools selection is started from easier data sources and tools. It is also a common expectation that the legal department will make the complexity arising from privacy requirements disappear and organize access to the confidential data. Either by initiating changing the laws in the country or negotiating contracts to get the confidential data without us changing our processes. It is too late when we discover, that the toolchain we created for working with open data is not capable of supporting confidentiality and privacy requirements or that the general public's perception of privacy intrusion is triggered by the statistics organization processing an innovative data source without corresponding innovative data protection measures.

4. **Rebuilding a statistical workflow with better privacy**

Often we find ourselves in the situation where a solution already exists in the organization, but the evolution of the external or internal environment demands a significant improvement in privacy guarantees. We will illustrate this using a statistics production process that uses mobile big data. The longest running official statistics application using mobile location big data is used by the Eesti Pank, the central bank of Estonia since 2009 [ref 5,6]. Eesti Pank, the second official statistics organization in the country compiles international travel statistics to estimate the current account of the balance of payments and calculate the import and export volumes of travel services. The statistics are calculated using data from mobile network operators. Up to date quarterly and annual statistics by countries is available to the public from the Eesti Pank website, see Figure 1.[1]

[1] http://statistika.eestipank.ee/#/en/p/1410/r/1770/1619

Figure 1



The developer of the system is an expert organization Positium LBS[2], concentrating on research and development of services using mobile location data. To improve the level of privacy protection and develop a future proofed version of the travel statistics software, Positium approached an R&D organization Cybernetica, specializing in privacy enhancing technology and services. Research and development started jointly with Cybernetica since 2016, before the GDPR came into force. It started with a privacy risk assessment, selection of candidate privacy enhancing technologies able to meet the functional needs of producing statistics on big data and an in-house legal evaluation of the planned solution. One of the requirements guiding the technology selection was cloud readiness to speed up deployment of future systems. The privacy enhancing technology selected was Sharemind HI framework for developing services with end-to-end data protection [ref 7]. In order to validate the technology a minimalistic prototype was build that captured the data and aggregated it into the analysis of trips. After the successful prototype, the actual design and development of the solution started. The selected privacy enhancing technology was explained to the national Data Protection Agency to prepare soliciting an official opinion of the DPA and to the national statistics organization to inform about the system with stronger technical privacy guarantees. Currently the

[2] http://positium.com/

new system is ready for producing statistics of inbound and outbound tourism and deployable into cloud environments supporting the required special hardware.

To our knowledge it is the only existing tourism statistics system with end-to-end data protection and auditability, ready to be deployed in countries respecting privacy of citizens and needing a cost-efficient solution based on mobile big data.

5. Self-assessment of Statistics Estonia against PbD: The 7 Foundational Principles

Privacy by Design: The 7 Foundational Principles by Ann Cavoukian [ref 8] are the general framework against which organization can assess their privacy management readiness. Statistics Estonia assessed their current situation concerning PbD and articulated where it hopes to be in 4 following years.

Simple qualitative scale on five categories was used in estimation. The categories were as follows: Very Good, Good, Acceptable, Poor, Very Poor. Estimation was given by experts of Statistics Estonia by their overall perception of the situation by analysing descriptions of the 7 foundational principles.

Short description was added to each principle explaining current situation (as-is) and perspective (to-be).

		Quality
1	Proactive not Reactive; Preventative not Remedial	Good
2	Privacy as the Default Setting	Acceptable
3	Privacy Embedded into Design	Acceptable
4	Full Functionality — Positive-Sum , not Zero-Sum	Poor
5	End-to-End Security — Full Lifecycle Protection	Poor
6	Visibility and Transparency — Keep it Open	Acceptable
7	Respect for User Privacy — Keep it User-Centric	Acceptable

1. Proactive not Reactive; Preventative not Remedial

Situation AS-IS: Overall estimation concerning proactivity and preventivness of the data sources already in use is good. One can't say the same about new type of sources, which more or less might be described as big data sources. Statistical confidentiality is followed in all cases.

Situation TO-BE: Good. Quality level must remain good, albeit substantive change in data sources, which turned to be new by type like sensor data, GPS data etc.

2. Privacy as the Default Setting

Situation AS-IS: Statistical confidentiality is followed in all cases. Automatic protection is not in place in all cases. In some data transfers and transformation protection is done by human intervention, which is lot of work.

Situation TO-BE: Good or very good. New PbD methods must ensure automatically protected privacy.

3. Privacy Embedded into Design

Situation AS-IS: In general architecture of Estonian Statistics information systems is designed taking into consideration privacy as core functionality. Still there is no clear where and how to make distinction between privacy and security issues. Risk of the offence to the privacy is higher in data catering and submission. PbD in information systems managing these functions must be improved.
Situation TO-BE: Good. Privacy by Design is embedded into all IT systems and business practices Estonian Statistics use.

4. Full Functionality — Positive-Sum, not Zero-Sum

Situation AS-IS: In current situation concern about security has been in focus. Persons as statistical units have no possibility to find out what kind of data there is in the holding of Statistics Estonia. There might be need to re-focusing in this point to get positive-sum "win-win" situation
Situation TO-BE: Acceptable. Dichotomy of privacy vs. security is better addressed. Statistical units have better overview about data governance concerning individual data about them.

5. End-to-End Security — Full Lifecycle Protection

Situation AS-IS: Data governance life cycle is definitely one aspect need to be improved. Secure retention is in general poorly managed both in the case of archiving including transfer to the National Archives and final disposition and delete of data.
Situation TO-BE: Acceptable. Data life cycle management is in place. Data governance fo which it is a part is implemented.

6. Visibility and Transparency — Keep it Open

Situation AS-IS: Currently there is absence of data stewards, who might act as data owners (stakeholders). Same is situation with data lineage, both business practice and technology need to be approved to present transparent data lineage, good governance and improve trust. There is no independent verification in place.
Situation TO-BE: Good or very good. Data stewards will be arrange to do both to govern data lineage and to explain it for trust and transparency for data users.

7. Respect for User Privacy — Keep it User-Centric

Situation AS-IS: This user-centric point is difficult to address in the official statistics framework. Official statistics has an aim to be user-centric, but disseminating product are aggregated data and indicators, not individual data. Largest impact is to the state registers, where Statistics Estonia might propose amendments. Sometime this adds characteristics or proposing use of classification and might be estimated as not user-friendly behaviour.
Situation TO-BE: Good. Where individuals have to provide data information systems will be designed by taking into account user-friendly and user-centric approach.

6. How to approach the change

In order for the European statistical system to evolve to take use of the opportunities from new data sources national and transnational producers of official statistics in Europe need to share best practices of including PbD in procurement with each other, educate staff in privacy management, conduct privacy impact assessments of future and existing statistical systems together with internal and external experts, update organizational policies and processes and finally update the national legal systems, making PbD the norm in development of European systems for official statistics.

To help, there is a growing list of international standards and ongoing standardization supporting operationalizing PbD principles [ref 9] while conducting privacy impact assessments, planning cloud computing, big data processing or considering privacy in IoT.
For instance ISO/IEC 29101:2013 - the standard on Privacy Architecture Frameworks brings architecture examples for information systems processing personal data and shows how privacy enhancing techniques: pseudonymisation, secure computing, query restrictions etc. can be deployed to protect personal information.
PbD becoming the norm in official statistics will give the citizens of Europe the insurance that the statistical system can move into the evolving area of trusted smart statistics and establish the trust of owners of evolving data sources helpful in producing official statistics.

7. Conclusion

Designing privacy into the systems from the start is plain good management and economical use of tax payer resources. Omitting PbD may result in scandals, loss of face for the statistics organization, loss of trust in the producer of official statistics and potentially cost managers' of the organization their job.
It is possible to create privacy-protecting alternatives to already existing solutions in order to prepare statistical organizations for handling sensitive big data. This usually needs the PbD approach and selection of privacy enhancing technologies.

8. References

1. Data Age 2025: The Evolution of Data to Life-Critical, Reinsel D, Gantz J, Rydning J, IDC 2017 https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

2. An Introduction to Privacy Engineering and Risk Management in Federal Systems, NISTIR 8062, https://doi.org/10.6028/NIST.IR.8062

3. Fundamental Principles of Official Statistics, 2013/21, UN Economic and Social Council https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf

4. EUROPEAN STATISTICS CODE OF PRACTICE For the National Statistical Authorities and Eurostat (EU statistical authority) https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf

5. Methodology for the compilation of international travel statistics http://statistika.eestipank.ee/failid/mbo/valisreisid_eng.html

6. Eesti Pank: Statistical indicators, International travel statistics http://statistika.eestipank.ee/#/en/p/MAKSEBIL_JA_INVPOS/1410

7. Sharemind HI for developing services with end-to-end data protection https://sharemind.cyber.ee/sharemind-hi/

8. Cavoukian, Ann (2009). "Privacy by Design: The 7 Foundational Principles." Office of the Information and Privacy Commissioner of Ontario.

9. Activities related to privacy standards in ISO https://ipen.trialog.com/wiki/ISO

Appendix A

Original description of PbD principles [ref 8]

1. Proactive not Reactive; Preventative not Remedial

The Privacy by Design (PbD) approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy invasive events before they happen. PbD does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred — it aims to prevent them from occurring. In short, Privacy by Design comes before-the-fact, not after.

2. Privacy as the Default Setting

We can all be certain of one thing — the default rules! Privacy by Design seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy — it is built into the system, by default.

3. Privacy Embedded into Design

Privacy by Design is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

4. Full Functionality — Positive-Sum , not Zero-Sum

Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. Privacy by Design avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

5. End-to-End Security — Full Lifecycle Protection

Privacy by Design, having been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the data involved — strong security measures are essential to privacy, from start to finish. This ensures that all data are securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, Privacy by Design ensures cradle to grave, secure lifecycle management of information, end-to-end.

6. Visibility and Transparency — Keep it Open

Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to

independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember, trust but verify

7. Respect for User Privacy — Keep it User-Centric

Above all, Privacy by Design requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric