

Privacy Impact Assessment of the 1994 US Data Census dataset

João Freitas up202100373

Rui Gonçalves up202103077

The present report, written in the context of PET (Privacy Enhancing Technologies) curricular unit at FCUP (Faculdade de Ciências da Universidade do Porto), aims at describing the process of assessing the privacy impact and risk of a dataset. It will start by performing a formal description of the methodology being applied to assess the dataset privacy impact. Secondly, the described methodology will be applied for the dataset, yielding privacy impact and risks scores and a set of actions that should be executed in order to lower these scores, followed by an analysis and evaluation of these. To conclude, an overview of the overall effectiveness and complexity of performing such assessment will be presented, as well as some remarks for future work.

1. Introduction to Privacy Impact Assessment

Defined by NIST (National Institute of Standards and Technology) [1] in the SP 800-53 publication [2], Privacy Impact Assessment (PIA) is the process of analyzing the risks and procedures used to work with Personal Identifying Information (PII), as a manner to evaluate and grant that the defined privacy requirements are complied (e.g., GDPR), as well as a way to identify associated privacy risks and measures to mitigate these [3]. Such process should be done before attempting to anonymize a dataset, as the risk and impact information that it reveals is critical for the context of the anonymization.

1.1. PIA Process

ISO 29134 provides a good description of the overall PIA process, as illustrated in Figure 1. Although this report has bigger focus on the **Perform PIA** stage, it is necessary to know the actions that occur before the PIA report, as well as the events that occur after.

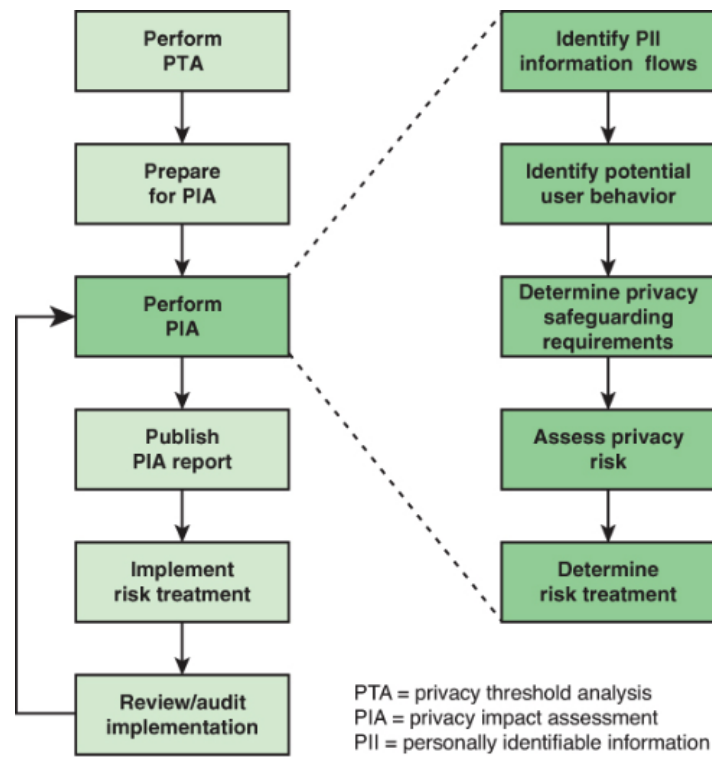


Figure 1 - PIA Process as defined in ISO 29134, source: [3]

In the first stage a **Privacy Threshold Analysis (PTA)** is conducted, which determines if whether or not a system requires a PIA. For this, a set of basic questions that relate to PII are proposed to the system owners [3]. If the system really requires a PIA, then the PIA starts being prepared by the PIA leader. The goal of this second stage is to create a strategic plan on how the PIA will occur, by identifying and forming the team that will execute it, describing the system in context and identifying the stakeholders of the system and those who are affected by it [3].

The third stage is the PIA report itself. As denoted in Figure 1 this stage can be broken into five small stages:

- **Identify PII Primary Flows**, where typically a workflow/swimlane diagram is draw in order to understand where the PII moves within and out of the system, for the collect, store, use, transfer and delete actions [3];
- **Identify Potential User Behavior**, which focuses on identifying specific user actions that may impact his privacy (e.g., accessing system through unprotected or shared Wi-Fi network) [3], [4];
- **Determine Privacy Safeguarding Requirements**, which focuses on the process of elicitation of national and international laws and regulations which restrict PII access (e.g., GDPR), current contracts that restrict the way PII is processed, business operations that require usage of PII and which Fair Information Practice Principles (FIPPs) [5] are relevant for the system [3]. By knowing the requirements which the system is imposed to, one can also assume the privacy controls that are required to be implemented;
- **Assess Privacy Risk**, which focuses on identifying the risk (i.e., potential threats and system vulnerabilities), analyzing and measuring the risk by calculating the impact levels of the threats, and evaluating the risk in order to know how to prioritize it [3];

- **Determine Risk Treatment**, which focuses on establishing solutions to mitigate the identified risks, by choosing treatment options, implement privacy and security controls and create risk treatment plans [3].

Beside PIA Report Publish, the remaining stages (Risk Treatment Implementation and Implementation Review/Audit) are not really part of the PIA process [3]. However, since the PIA is an iterative process, it makes sense to include them in the lifecycle of PIA [3].

1.2. Assessing Privacy Impact

Privacy Impact can be defined as the severity to which information has to the individual that it belongs and to the organization that detains it. NIST SP 800-53 refers that privacy impact can be estimated by two factors: prejudicial potential and level of identification [2]. The former relates to the damage that can be caused if the information privacy is breached and the latter refers to the difficulty to which the information can be used to identify an individual. Typically, these factors are categorized in five levels – **Very Low**, **Low**, **Moderate**, **High** and **Very High** – denoting different impact definitions for both individuals and organizations, as well as for the risk of identification, as described in Table 1 and Table 2.

Impact	Individuals	Organization
Very Low	Virtually no noticeable impact	Virtually no noticeable impact
Low	Negligible economic loss; or small, temporary reduction of reputation; or no impact on other personal factors	No violation of law or regulation; or negligible economic loss; or small, temporary reduction of reputation
Moderate	Economic loss that can be restored; or small reduction of reputation or at most minor impact on other personal factors	Minor violation of law or regulation resulting in warning; or economic loss that can be restored; or some reduction of reputation
High	Large economic loss that cannot be restored; or serious loss of reputation or other psychological damage by revealing sensitive information; or serious impact on other personal factors	Violation of law or regulation resulting in a fine or minor penalty; or large economic loss that cannot be restored; or serious and long-lasting loss of reputation
Very High	Considerable economic loss that cannot be restored; or serious loss of reputation or other psychological damage with long-lasting or permanent consequences; or serious impact on other personal factors	Serious violation of law or regulation resulting in a substantial fine or other penalty; or considerable economic loss that cannot be restored; or devastating and long-lasting loss of reputation

Table 1 – Impact levels definition for both individuals and organizations, adapted from: [3]

Identification Level	Definition
Very Low	Identifying an individual using his or her personal data appears to be virtually impossible
Low	Identifying an individual using his or her personal data appears to be difficult but is possible in certain cases
Moderate	Identifying an individual using his or her personal data appears to be of only moderate difficulty
High	Identifying an individual using his or her personal data appears to be relatively easy
Very High	Identifying an individual using his or her personal data appears to be extremely easy

Table 2 – Identification levels definitions, source: [3]

Having understood what privacy impact is and how it can be measured, one just needs to start contextualizing their system into these impact and risk levels. For that, the E.U. Smart Grid Task Force (SGTF18) describes a rather simple methodology, that guides organizations in measuring the systems these need to protect [3], [6]. It starts by requiring the identification of the threat categories that are relevant to the organization environment, and for each of these, the primary assets (existing PII that need protection) associated with each are also identified (1). Then, the primary assets are described in order to find the relevant damage points that may occur on asset breach, as well as the easiness of information identification, which yields the prejudicial effects and identification levels of the system (2). To conclude, some calculations to normalize the impact scores are performed, by summing assets identification and prejudicial effects levels and comparing these values with a normalization scale table, as the one proposed in Table 3 (3). As a final note, the impact of the threat categories identified previously is given by the maximum impact value found in the assets that correspond to such category.

Identification + Prejudicial Effects	<4	4-5	6	7	>7
Impact	Very Low	Low	Moderate	High	Very High

Table 3 – Proposed normalization scale by SGTF18, source: [6]

1.3. Assessing Privacy Risk

Privacy Risk can be defined as the probability which consequences may occur when processing PII of an individual, in a non-anonymized manner [3], [7], [8]. To measure privacy risk, it is first needed to identify the likelihood levels of privacy, by comparing the risk level of four different factors [3]:

- **Capability of Threat Source**, which relates the intentions and power which one has (human or not-human source) to carry out a threat;
- **Threat Event Frequency**, which relates the number of times that the threat source will attempt to carry out a threat, during a time period;
- **Vulnerability**, which measures how vulnerable and easy the system is to be breached;
- **Control Effectiveness**, that quantifies the probability of a threat to occur, by analyzing the security of the system.

Likelihood levels are identified in a similar way as privacy impact levels. First, for each likelihood factor it is identified the meaning of each level, based on the organization environment. Having a consensus on what each level represents for each factor, these can be scored in light of the threat category in context. Once the factors have been scored, two pairs of factors need to be defined, so that the sum of their scores can be compared with Table 3. This will output two normalized values, that when combined and normalized, results in the likelihood level for the threat category.

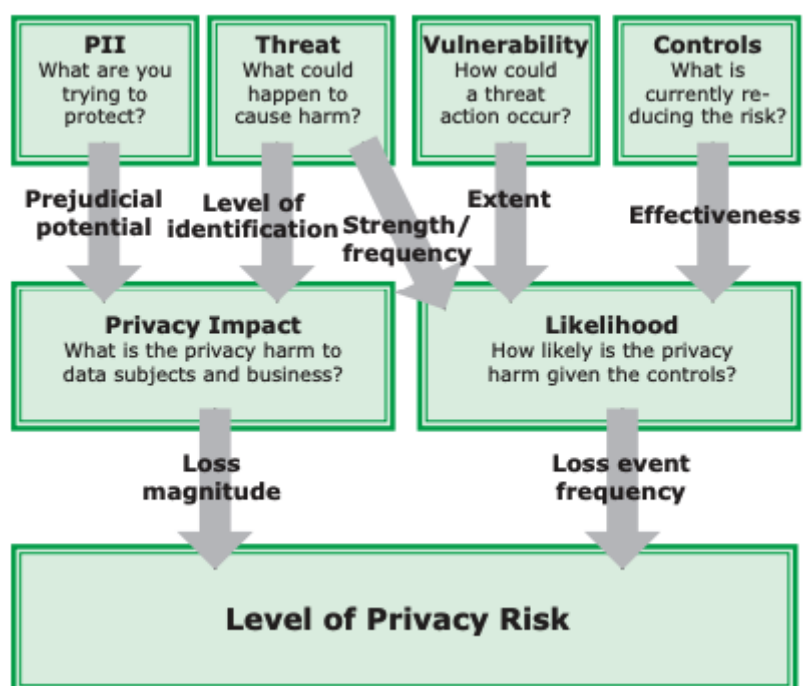


Figure 2 - Visual rundown of assessing privacy risk, source: [3], [4]

As described in Figure 2, having identified the likelihood levels of the system, these can be paired with the privacy impact scores, creating a 5x5 risk matrix similar to the one described in Figure 3. The level which results of the matching pair in the matrix identifies the overall risk of the threat category.

		Privacy impact				
Level of identification	VH	VH	VH	VH	VH	VH
	H	H	H	VH	VH	VH
	M	M	M	H	VH	VH
	L	L	M	H	VH	VH
	VL	L	M	H	VH	VH
		VL	L	M	H	VH
		Prejudicial potential				

Figure 3 - Example of a privacy risk matrix, source: [3]

The risk levels defined for the matrix need to be described based on the organization environment and the threat category context. This is required as the impact of a threat may be generally higher than the likelihood of it happening, and vice-versa. However, there are general guidelines that help defining these risks, such as the ones described in Appendix A.

2. Performing PIA on the Census Dataset

After understanding how the PIA process is executed, it is now possible to apply it on the dataset in context. As mentioned previously in Chapter 1, the main focus of this report is to perform a PIA report and as a result the first stage will be ignored as there is no background knowledge about the privacy leader that ordered the PIA. However, to get better results from the PIA report, a fictitious European organization that detains the dataset will be assumed. Additionally, the St. Anywhere's PIA report [9] will be used as a template as well as the open-source PIA software tool from CNIL (Commission, Nationale de l'Informatique et des Libertés) [10], as these two are powerful sources and tools for automating the PIA report development. While the present chapter presents a detailed summary of the PIA process conducted, the full output of the PIA software tool can be found in Appendix B.

2.1 Background and Introduction

Têrepê is a small Portuguese software house that mainly works on Data Science projects for European and American organizations, being constituted by two teams of 10 software engineers, 5 data scientists, 2 product owners, 3 UI designers and 5 board members (25 in total). The company is completely in-house and as so does not sub-contract any company in their projects. Their latest project is financed by the US Census Bureau [11] (\$2M) and involves creating a machine learning tool that given a subset of the Census, predicts whether or not an American Citizen has a yearly income bigger than \$50k. Knowing that the US Census Bureau is part of the US government and that there has been an increase in user privacy protection, Têrepê CISO (Chief Information Security Officer) and DPO (Data Protection Officer) gathered together to form a task-force, in order to make sure that every law and security aspects were being comply. They have then decided to create a PIA for the project.

There is only one team (10 members) working for the census project and the Bureau specifically assigned them the 1994 US Data Census dataset. Census is a form of activity that aims to inquire the citizens of a country, in order to get insights on the quality of life these are having, as well as their pain-points [12]. This way, the government of the inquired country can get feedback from their citizens and in return improve their lives by investing in the country infrastructure [12]. In the case of the US Census, these are conducted every 10 years [13].

Although not including all information of the Census, the dataset encompasses a large range of attributes (i.e., data identifiers) (14) that characterize an individual biological and identity status (age, sex, race, native country, relationship, marital status), social status and education (work class, occupation, weekly work hours, education), net worth (capital gain, capital loss), as well as some metadata attributes that do not entirely identify an individual. The dataset only contains records of individuals whose age is larger than 16.

2.2 Stage 2 – Prepare for PIA

As previously described, the US Census Bureau contacted Têrepê for the production of the predictive tool that uses as input the US Census, with a budget of \$2M. The team allocated for the project is constituted by 5 software engineers, 3 data scientists, 1 product owner and 1 UI designer. They are responsible for developing and maintaining the project for a span of 2 years. The team data scientists are the only ones who can have direct contact with the provided dataset, while the software engineers are only aware of the attributes that it contains. The product owner and UI designer are responsible for creating the most value for the tool and have weekly meetings with the data scientists, Têrepê board members and Bureau stakeholders.

The company CISO and DPO are the main stakeholders for the project security and privacy, and their responsibilities are to discover every present and future privacy concern that might be considered a threat for the user privacy. While the CISO focus more the security aspect of the tool and company, the DPO is continuously finding ways in which the project might be violating any law or policy specified in the GDPR (General Data Protection Regulation) [14], E-Government Act of 2002 [15], CalOPPA (California Privacy Protection Act) [16] and CCPA (California Consumer Privacy Act) [17].

The PII processed by the predictive and analytic tools characterize an individual biological and identity status (age, sex, race, native country, relationship, marital status), social status and education (work class, occupation, weekly work hours, education), net worth (capital gain, capital loss). There are two versions of the dataset held by the company: the original dataset, that is kept on a hard drive that was flown by an air courier by the Bureau, stored in a secure vault inside the company; and a copy of the dataset in an anonymized form, stored in the company on-premises servers. These two datasets are only held in the company for the span of the project period.

Regarding the tools, platforms and technologies which the team can work it, the following have been admitted by the CISO:

- Windows 10, version 1909;
- Ubuntu x64 20.04;
- Visual Studio Code, version 1.63;
- Github Enterprise 2021;
- Microsoft Office Tools (Outlook, Word, Powerpoint, Excel, Teams) 2021;
- Redis Database, version 6.0.9;
- PostgreSQL, version 13.4;
- OpenVPN, version 2.5.5.

Any software that teams members use that is not present in the supported list is not covered by the company and thus should be the employee responsibility to pay for damages caused, if any issue occurs. The company also does not enforce the use of any specific programming languages, as these are to be decided by the team working on the projects.

2.3 Stage 3 – PIA Perform

Before identifying and assessing the risks associated with the project, it is necessary to identify the primary PII flows, in order to gather and understand the operations that affect its privacy. Figure 4 represents these using a swimlane diagram.

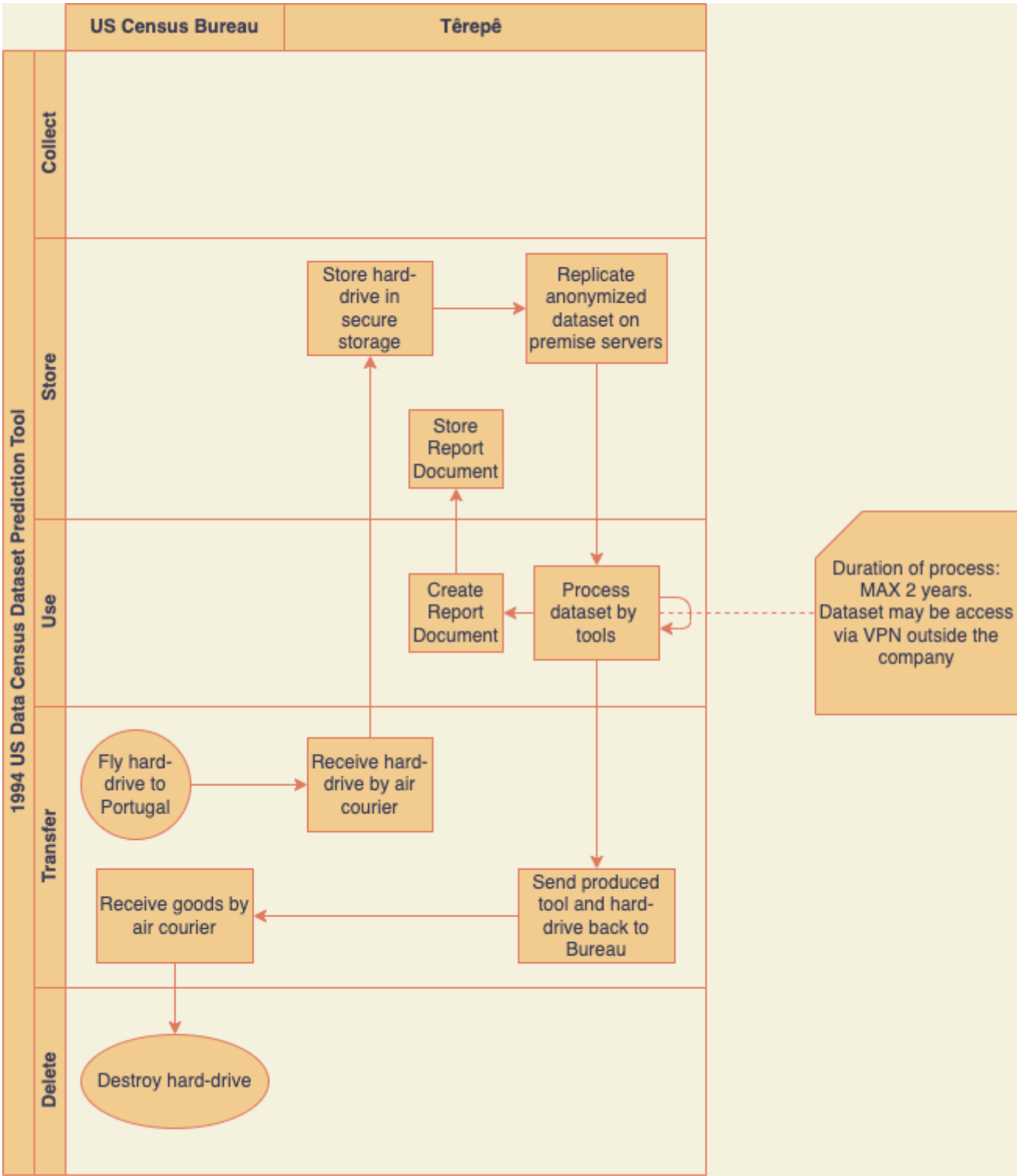


Figure 4 - Lifecycle of the dataset PII, represented in a swimlane diagram

No direct identifiers are present in the dataset, so data scientists cannot indicate which US citizen an entry in the dataset corresponds to. The dataset does have indirect and sensitive identifiers that can be used for data correlation and thus predict which US Citizen dataset entry corresponds to. To grant that only data scientists are allowed to access the data, a dedicated server to access the data was created, which only the data scientists have the keys to access it. The data of the dataset is

also considered accurate and up to date as it was provided by the US Census Bureau. Beside outliers and unnecessary data removal, the dataset is not modified.

2.3.1 Fundamental Principles Controls to Protect the Personal Rights of Data Subjects

Regarding the control to protect the personal rights of data subjects, the US Census Bureau has made a public announcement on their website stating that the census data will be used for learning about the quality of life of citizens [12]. There is no possibility to request data removal from the census as filling it is mandatory for all US Citizens. This means that no data subject can request data requests to Têrepê. For online privacy data requests in the Census Bureau website, US Citizens can contact for the Chief FOIA Officer at 1-800-432-1494, or by email at pco.policy.office@census.gov [18].

The only processing and storage operations performed by Têrepê are conducted in Portugal. Operations conducted in the US or outside the European Union by the US Census Bureau is not a concern of Têrepê. Additionally, the Bureau signed an NDA contract with all Têrepê employees, stating that for a period of 5 years, these could not reveal that they are working with the US Data Census or with the US Census Bureau, neither disclose information about the Census.

2.3.2 Measures of Risk Control

Taking in consideration the data operations and security measures described, the following risk measures have been identified:

Encryption

At Têrepê offices, Internet access is barred by a secure firewall that makes sure to encrypt any unencrypted connections with TLS 1.1 and 1.2. Outside Têrepê offices, access to the company infrastructure is done by using a VPN using the OpenVPN software. The VPN tunnel is protected using IPsec IKE v2. There is still a risk of disclosure of information by employees when accessing outside the company in unprotected Wi-Fi networks.

Anonymization

The US Data Census dataset has been anonymized before processing by analytic tools using the ARX software tool. Although anonymized, there is still a risk of information disclosure if it revealed to an attacker, as the anonymization had to take into account the needed utility levels for the prediction tool to yield good results.

Logical Access Control

Data scientists are the only employees who have access to the dataset. A dedicated server to access the dataset was created which can only be accessed using keys distributed to these employees in the form of USB Yubi Keys [19]. There is still a risk of gaining unauthorized access by internal employees or external attackers, if someone manages to grab/steal the USB Yubi Key.

Paper Document Security

Product Owners and UI Designers do weekly reports on the tool being created in order to refine its features. For this they use Microsoft Word to create a report, which is then translated in the .PDF format and distributed to board members for presentations and validations. Every document is stored in a secure storage and destroyed after two years.

There is still a risk of disclosure of information by employees if they share these report documents with those that are not on the meetings, either via Social Engineering attacks in E-mails or by talking about it.

Network Security

CISO implemented and deployed a network topology that involves an IDS, a DMZ and firewalls to secure and detect unauthorized accesses. There is still a risk of disclosure of information if employee install malware or a compromised software.

2.3.3 Risk Assessment

To assess the project risk, it is first necessary to understand the impacts and threats for the risk of illegitimate access to data, unwanted modification of data and data disappearance. A summary of these can be found below, in Table 4, which also includes the identified sources and measures to address each risk.

Risk	Impacts	Threats	Sources	Measures
Illegitimate Access to Data	Breach of Net Worth and Identity, Target of Marketing Companies and Robbers, Public Damage and Political and Social-Economical Advantage to Rival Countries	Social Engineering, Packet Sniffing, Supply-Chain Attacks, Malware, Proprietary Device Thief, Human Interaction	Employees, Software	Encryption, Anonymization, Logical Access Control, Paper Document Security, Network Security
Unwanted Modification of Data	Public Damage	Malware, Proprietary Device Thief, Human Interaction	Employees	Logical Access Control

Risk	Impacts	Threats	Sources	Measures
Data Disappearance	Public Damage	Social Engineering, Packet Sniffing, Supply-Chain Attacks, Malware, Proprietary Device Thief, Human Interaction	Employees, Software	Encryption, Logical Access Control, Network security

Table 4 – Impacts, Threats, Sources and Control Measures for identified risks

For the score classification of the impact and likelihood of risks, the following scale was used:

Undefined	Negligible	Limited	Important	Maximum
1	2	3	4	5

Table 5 – Scale used to classify impact and likelihood scores of risks

The Illegitimate Access to Data was considered to be the most severe risk (**Maximum** impact and **Important** likelihood) as it implies the breach of the dataset. Breaching citizens data of the world's most powerful country [20] would be catastrophic, not only because it would put the country citizens identity and status at risk, but also because it would cause enormous fines to the company, Portugal and the United States. Additionally, it would give political advantage to rival countries such as China and Russia. The likelihood is lower than the impact as the company data processing is well secured and the dataset does not direct identifiers, meaning that it would require secret datasets from the US to correlate the data entries.

Unwanted Modification of Data risk is not impactful (**Negligible** impact and likelihood), as it does not matter if the employees at Têrepê modify the data, as they are creating a predictive tool that is agnostic of the content of the dataset. The US Census Bureau will use whatever dataset they desire from the census.

Finally, the Data Disappearance risk is impactful, but not for the primary data subjects of the dataset. Since the US Census Bureau detains the primary source of the dataset, losing the hard drive would not mean that the citizens data is lost. However, it would be really bad for the company to lose the dataset, as it would leave a really bad impression for the US Government. It would have a serious impact if that data would be lost in terms of a robbery, but that falls under the Illegitimate Access to Data risk.

A visual overview of the risks impact can be found below in Figure 5, as well as how the risks are associated to the identified impacts, threats, sources and measures in Appendix B.

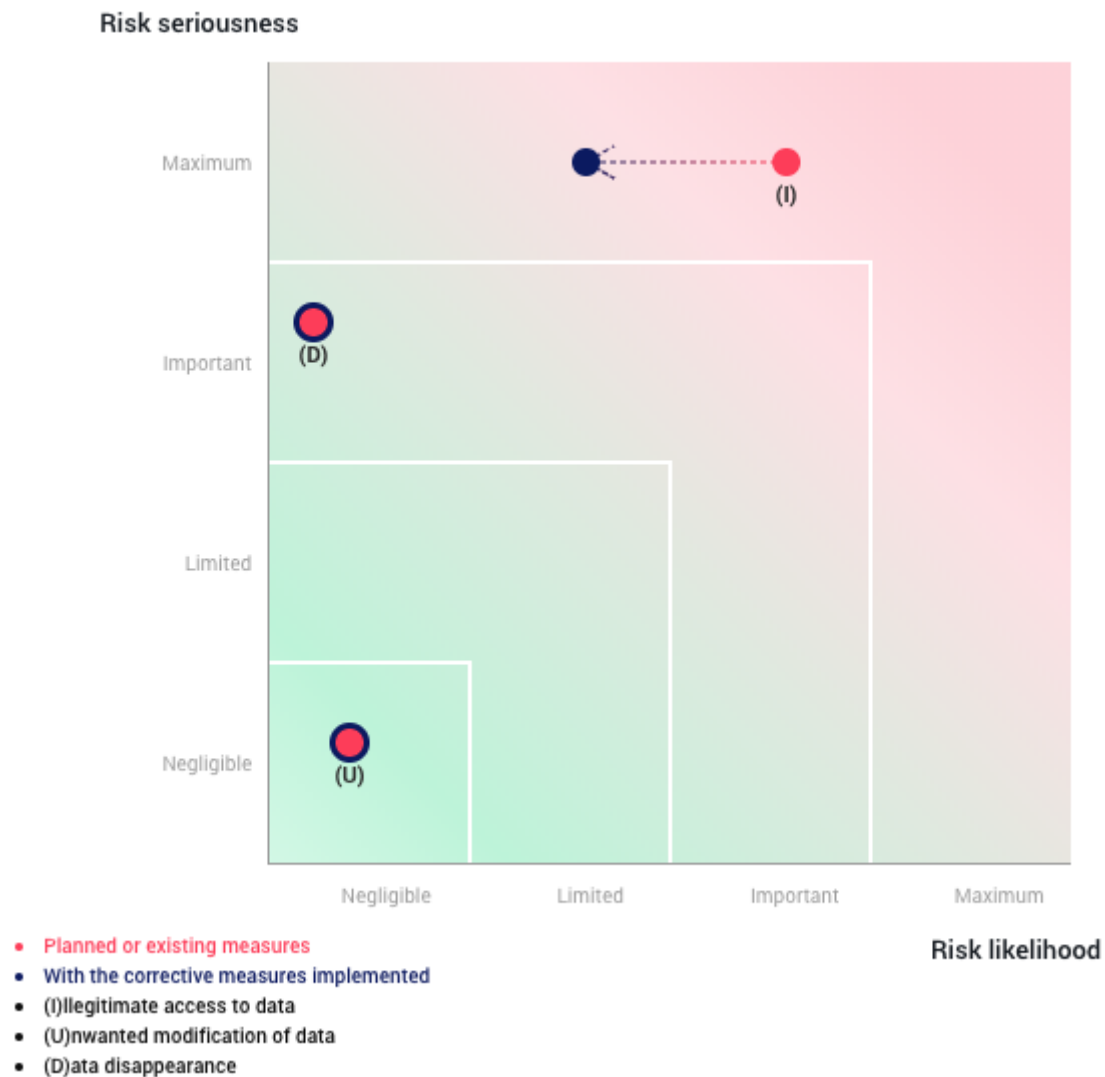


Figure 5 - Risks impact and likelihood represented in a Risk Matrix

2.3.4 Action Plan

Previously, in Section 2.3.2, it has been identified a set of five risk control measures that allow to address the privacy and security risks. Despite these helping to lower the likelihood of the risks, it has also been mentioned that these have some inherent risks regarding information disclosure. To address these risks some additional actions have been planned, as well as the responsible entities and deadlines to implement these, which can be found below, in Table 6.

Measure	Action	Responsible Entities	Deadline
Anonymization	Re-anonymize the dataset to grant that the privacy levels are bigger than the utility levels, but at the same time the analytic tools can provide good results.	Data Scientists	31st January, 2022

Measure	Action	Responsible Entities	Deadline
Logical Access Control	Deploy a GPS security mechanism that allows the track of the USB YUBI keys. This will reduce the likelihood of the loss of proprietary devices.	CISO, CEO	31st January, 2022
Paper Document Security	Deploy an IDS rule to block unrecognized e-mail domains.	CISO	31st January, 2022
Network Security	Setup anti install mechanism on employees' computers, to prevent install of malware.	Data Scientists	31st January, 2022

Table 6 – Action plan defined for control measures

The CISO and DPO believe that implementing these actions will reduce the likelihood of the risks, especially for the Network Security and Paper Document Security measures, as it should be impossible to install external Software without CISO recognition and receive e-mails from external domains. However, and despite lowering the likelihood, there is no guarantee that data scientists USB YUBI keys will not be stolen by an attacker. Increasing the privacy level of the dataset anonymization will also not guarantee that it is impossible to correlate with external datasets, as there should be an optimal level of utility for the analytic tools be able to process the dataset.

In addition to the action plans proposed for the risk measures, after implementing these it will also be possible to reduce the impact for the Illegitimate Access of Data and Data Disappearance risks. Figure 6 summarizes the action plans proposed in the PIA.

Overview

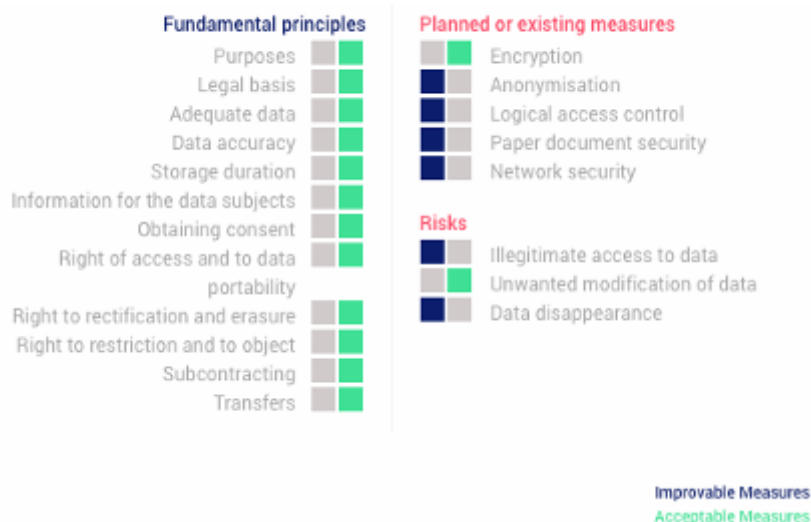


Figure 6 – Overview of the proposed PIA action plan. Blue boxes indicate improvements

3. Conclusions and Future Work

At a first glance the PIA process methodology seems too much complex, verbose and very time consuming. However, using software specialized software such as the open-source CNIL PIA software, the complexity associated gets lower and lower, it is just necessary to know the background and context of the data subjects and fill the boxes with the requested information. A PIA also helps finding concerns and constraints which have not been assumed previously. However, to perform a good PIA, it is necessary to train companies employees on data law and computer security as the roles of those who perform and evaluate the PIA is critique.

For future works it would be nice to perform the PIA in a more real environment, since it was required to make a lot of assumptions regarding what the company was provided and what not. Additionally, measures should be quantified in a way that action plans can be refined even more.

4. References

- [1] NIST, 'National Institute of Standards and Technology', NIST, 2021. <https://www.nist.gov/> (accessed Dec. 28, 2021).
- [2] Joint Task Force Interagency Working Group, 'Security and Privacy Controls for Information Systems and Organizations', National Institute of Standards and Technology, Sep. 2020. doi: 10.6028/NIST.SP.800-53r5.
- [3] W. Stallings, Information Privacy Engineering and Privacy by Design: Understanding privacy threats, technologies, and regulations, Original retail. 2020.
- [4] W. Stallings, 'Privacy Impact Assessment: The Foundation for Managing Privacy Risk', Priv. Impact Assess., p. 13, Mar. 2021.
- [5] Cloudflare, 'What are the Fair Information Practices? | FIPPs', Cloudflare, 2021. <https://www.cloudflare.com/learning/privacy/what-are-fair-information-practices-fipps/> (accessed Dec. 30, 2021).
- [6] E.U. Smart Grid Task Force, 'Data protection impact assessment template for smart grid and metering systems', Energy - European Commission, Sep. 27, 2018. https://ec.europa.eu/energy/content/data-protection-impact-assessment-template-smart-grid-and-smart-metering-systems_pt (accessed Dec. 30, 2021).
- [7] C. C. NIST, 'Privacy Risk - Glossary | CSRC', 2021. https://csrc.nist.gov/glossary/term/privacy_risk (accessed Dec. 30, 2021).
- [8] S. Brooks, M. Garcia, N. Lefkovitz, S. Lightman, and E. Nadeau, 'An introduction to privacy engineering and risk management in federal systems', National Institute of Standards and Technology, Gaithersburg, MD, NIST IR 8062, Jan. 2017. doi: 10.6028/NIST.IR.8062.
- [9] HIQA, 'Sample Privacy Impact Assessment Report Project: Outsourcing clinical audit to an external company in St. Anywhere's hospital', Oct. 2010. [Online]. Available: https://www.hiqa.ie/sites/default/files/2017-02/HI_PIA_Sample_Report.pdf
- [10] CNIL, 'The open source PIA software helps to carry out data protection impact assesment | CNIL', Jun. 30, 2021. <https://www.cnil.fr/en/open-source-pia-software-helps-carry-out-data-protection-impact-assesment> (accessed Dec. 31, 2021).
- [11] U. C. Bureau, 'About the Bureau', [Census.gov](https://www.census.gov/), 2021. <https://www.census.gov/about-us> (accessed Dec. 31, 2021).
- [12] U. C. Bureau, 'What We Do', [Census.gov](https://www.census.gov/), Nov. 18, 2021. <https://www.census.gov/about/what.html> (accessed Dec. 05, 2021).

- [13] GOV.UK, CENSUS, 'About the census', Census 2021, 2021. <https://census.gov.uk/about-the-census> (accessed Dec. 01, 2021).
- [14] GDPR, 'General Data Protection Regulation (GDPR) – Official Legal Text', General Data Protection Regulation (GDPR), 2021. <https://gdpr-info.eu/> (accessed Jan. 01, 2022).
- [15] N. Archives, 'E-Government Act of 2002', National Archives, Aug. 15, 2016. <https://www.archives.gov/about/laws/egov-act-section-207.html> (accessed Jan. 01, 2022).
- [16] P. Policies, 'CalOPPA: California Online Privacy Protection Act', Privacy Policies, Nov. 15, 2021. <https://www.privacypolicies.com/blog/caloppa/> (accessed Jan. 01, 2022).
- [17] R. Bonta, 'California Consumer Privacy Act (CCPA)', State of California - Department of Justice - Office of the Attorney General, Oct. 15, 2018. <https://oag.ca.gov/privacy/ccpa> (accessed Jan. 01, 2022).
- [18] U. C. Bureau, 'Online Privacy Policy', [Census.gov](https://www.census.gov), 2021. <https://www.census.gov/about/policies/privacy/privacy-policy.html> (accessed Jan. 01, 2022).
- [19] Yubico, 'Home', Yubico, 2021. <https://www.yubico.com/> (accessed Jan. 01, 2022).
- [20] WorldPopulationReview, 'Most Powerful Countries 2021', 2021. <https://worldpopulationreview.com/country-rankings/most-powerful-countries> (accessed Jan. 01, 2022).

Appendixes

A – General Rules for Assessing Privacy Risk Levels (Adapted: [3], [4])

Very High: These risks must be absolutely avoided or significantly reduced by implementing controls that reduce both their impact and likelihood [6] recommends that the organization implement independent controls of prevention (actions taken prior to a damaging event), protection (actions taken during a damaging event) and recovery (actions taken after a damaging event).

High: These risks should be avoided or reduced by implementing controls that reduce the impact and/or likelihood, as appropriate. The emphasis for these risks should be on prevention if the impact is relatively high and the likelihood is relatively low, and on recovery if the impact is relatively low and the likelihood relatively high.

Moderate: The approach for moderate risk is essentially the same as for high risk. The difference is that moderate risks are of lesser priority and the organization may choose to devote less resources to addressing them.

Low: The organization may be willing to accept these risks without further control implementation, especially if the treatment of other security or privacy risks also reduce this risk.

Very low: The organization may be willing to accept these risks because further attempts at reduction are not cost effective.

Preview

GENERAL INFORMATION

Preview

Editing :

Evaluation :

Validation :


João, Freitas

Rui, Gonçalves

João, Vilela

Status :

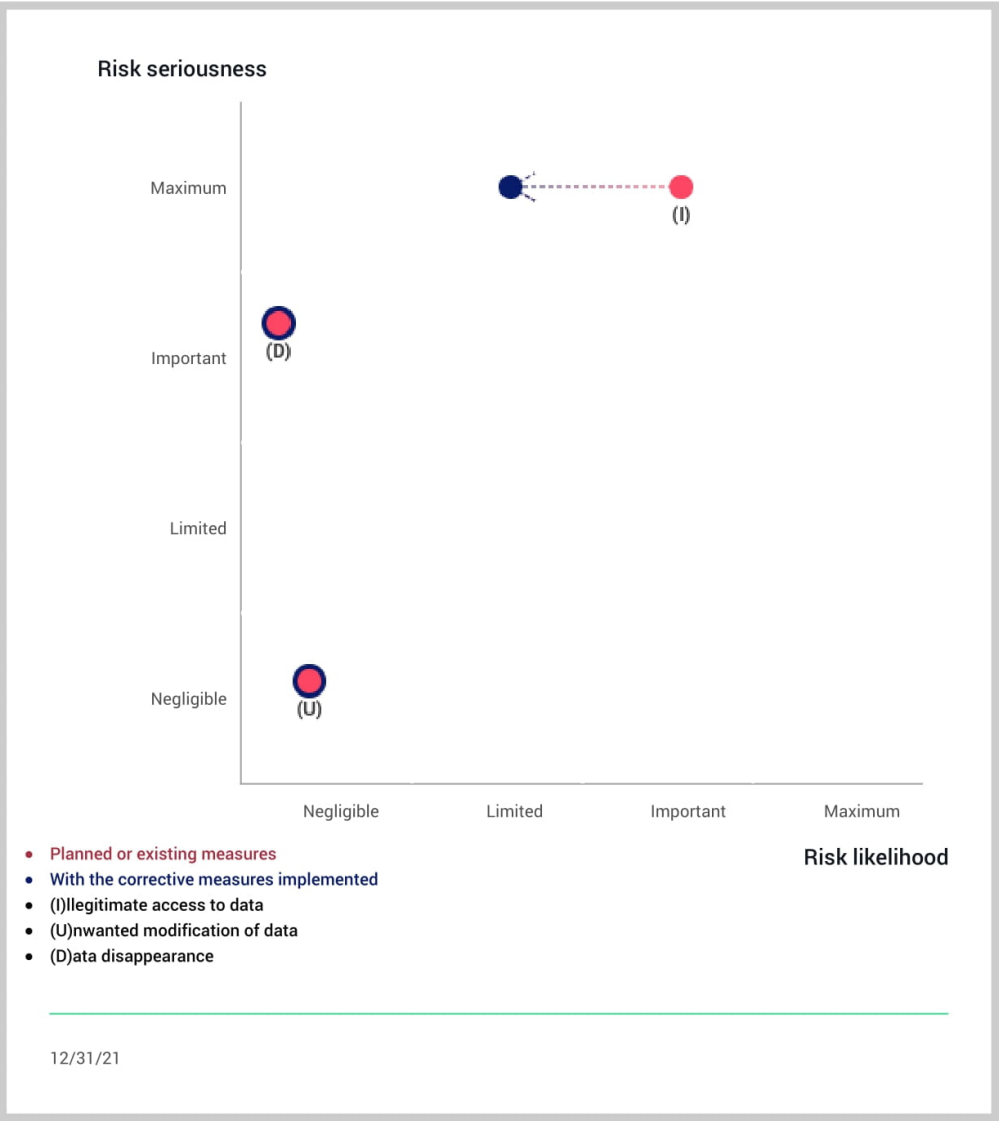
Simple validation

edit

100%

Validation

Risk mapping



Validation

Action plan

Fundamental principles	Existing or planned measures
<ul style="list-style-type: none"> Purposes Legal basis Adequate data Data accuracy Storage duration Information for the data subjects Obtaining consent Right of access and to data portability Right to rectification and erasure Right to restriction and to object Subcontracting Transfers 	<ul style="list-style-type: none"> Encryption Anonymisation Logical access control Paper document security Network security
	Risks <ul style="list-style-type: none"> Illegitimate access to data Unwanted modification of data Data disappearance
	Improvable Measures Acceptable Measures

Fundamental principles

No action plan recorded.

Existing or planned measures

Anonymisation

Action plan / corrective actions :

Re-anonymize the dataset to grant that the privacy levels are bigger than the utility levels, but at the same time the analytic tools can provide good results.

Evaluation comment :

Although anonymized, there is still a risk of information disclosure if it revealed to an attacker, as the anonymization had to take into account the needed utility levels for the prediction tool to yield good results.

Expected date of implementation : 1/31/22

Responsible for implementation : Data Scientists

Logical access control

Action plan / corrective actions :

Deploy GPS Security Mechanism to track USB Yubi keys of data scientists

Evaluation comment :

Deploy a GPS security mechanism that allows the track of the USB Yubi keys. This will reduce the likelihood of the lose of proprietary devices.

Expected date of implementation : 1/31/22

Responsible for implementation : CISQ, CEO

Paper document security

Action plan / corrective actions :

Deploy a IDS rule to block unrecognized e-mail domains.

Evaluation comment :

Deploy a IDS rule to block unrecognized e-mail domains. This will prevent social-engineering attacks by outsiders.

Expected date of implementation : 1/31/22

Responsible for implementation : CISO

Network security

Action plan / corrective actions :

Setup anti install mechanism on employees computers, to prevent install of malware.

Evaluation comment :

Setup anti install mechanism on employees computers, to prevent install of malware. Employees that need to install any software will have to check with the CISO before, so he can validate if it is malware or not.

Expected date of implementation : 1/31/22

Responsible for implementation : CISO

Risks - Illegitimate access to data

Action plan / corrective actions :

- Deploy GPS security mechanism to reduce likelihood of Proprietary Device Steal;
- Deploy IDS rule to prevent likelihood of Social Engineering via e-mail;
- Block installation of software that is not present on the supported software list;
- Re-anonymize dataset to increase privacy but reduce utility levels, as a measure to reduce the likelihood of correlating the anonymized dataset

Evaluation comment :

Reverted likelihood risk to Limited

Added action plan

Expected date of implementation : 2/28/22

Responsible for implementation : CISO, Data Scientists

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Illegitimate access to data)? **Maximum**

Taking into account the action plan, how do you re-evaluate the **likelihood of this risk** (Illegitimate access to data)? **Limited**

Risks - Data disappearance

Action plan / corrective actions :

- Improve hardware security mechanisms to protect access to external hard drive in the company facilities.

Evaluation comment :

Added action plan

Expected date of implementation : 2/28/22

Responsible for implementation : CISO

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Data disappearance)? **Important**

Taking into account the action plan, how do you re-evaluate the **likelihood of this risk** (Data disappearance)? **Negligible**

Validation

TO TRANSLATE - DPO and data subjects opinion

DPO's name

João Freitas

DPO's status

The treatment could be implemented.

DPO's opinion

PIA is valid and actions plans were defined.

Search of concerned people opinion

Concerned people opinion was requested.

Concerned people opinions

US Census Bureau

Concerned people statuses

The treatment could be implemented.

Concerned people opinions

"Dataset security will increase with the proposed action plan, which means that the privacy of the dataset will too"

Context

Overview

What is the processing under consideration?

The 1994 US Data Census dataset is part of one of the machine learning projects of Têrepê, a Portuguese software house company. The purpose of the project is to create a tool that predicts whether or not an American citizen has an income bigger than \$50,000 per year. For that, the US Census Bureau provided the dataset for study.

The principal stateholders for the project is the US Census Bureau and Têrepê CISO and DPO.

What are the responsibilities linked to the processing?

The team allocated for the project is constituted by 5 software engineers, 3 data scientists, 1 product owner and 1 UI designer. They are responsible for developing and maintaining the project for a span of 3 years.

The CISO is responsible for the company overall information security. Together with DPO, they are finding every privacy concern a flaw which might be considered a threat for the user privacy.

Are there standards applicable to the processing?

Europe:

GDPR - General Data Protection Regulation

America:

California Privacy Protection Act (CalOPPA)

California Consumer Privacy Act (CCPA)

Evaluation : Acceptable

Context

Data, processes and supporting assets

What are the data processed?

The data processed characterize an individual biological and identity status (age, sex, race, native country, relationship, marital status), social status and education (work class, occupation, weekly work hours, education), net worth (capital gain, capital loss).

The original data was given to the company in an hard-drive, flown by the US Census Bureau. The company keeps the hard-drive in a secure storage inside the company and a copy of the dataset inside the company on-premise servers. Developers can access data via VPN in any location external to the company.

The data is to be stored for a span of 2 years (December 2021 - December 2023), which corresponds to the duration of the project.

How does the life cycle of data and processes work?

An overview of the project lifecycle can be found in the attachements in the form of a Swimlane diagram (trp-data-census-pii-primary-flows.png).

What are the data supporting assets?

The company only allows the use of the following technologies and platforms to work with:

- Windows 10, version 1909
- Ubuntu x64 20.04
- Visual Studio Code, version 1.63
- Github Enterprise 2021
- Microsoft Office Tools (Outlook, Word, Powerpoint, Excel, Teams) 2021
- Redis Database, version 6.0.9
- PostgreSQL, version 13.4
- OpenVPN, version 2.5.5

The company does not enforce the use of any programming languages as these are to be decided by the team working on the projects. Any software that team members use that is not present in the supported list is not covered by the company if any issue occurs and thus should be the worker responsibility to pay for damages caused.

Evaluation : Acceptable

Fundamental principles

Proportionality and necessity

Are the processing purposes specified, explicit and legitimate?

Processing of the dataset is required to create machine learning tools. For this, analytic tools ingest the dataset in order to classify it and create learning models for the prediction in context. Datascientists are the entities responsible to access the data, and use the analytic tools.

Evaluation : Acceptable

What are the legal basis making the processing lawful?

No direct identifiers are present in the dataset, so datascientists cannot indicate which US citizen an entry in the dataset corresponds to. The dataset does have indirect and sensitive identifiers that can be used for data correlation and thus predict which US Citizen dataset entry corresponds to. To grant that only datascientists are allowed to access the data, a dedicated server to access the data was created, which only the data scientists have the keys to access it.

Evaluation : Acceptable

Are the data collected adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')?

The majority dataset attributes are required for building the prediction tool as these relate to the individual capital power. Attributes fnlwgt, education-num can be removed from the dataset as these are metadata attributes that are not beneficial for the prediction.

Evaluation : Acceptable

Are the data accurate and kept up to date?

Data is considered accurate and up to date as it was provided by the US Census Bureau. Beside outliers and unnecessary data removal, the dataset is not modified.

Evaluation : Acceptable

What are the storage duration of the data?

Data is stored for the period of 2 years.

Evaluation : Acceptable

Fundamental principles

Controls to protect the personal rights of data subjects

How are the data subjects informed on the processing?

The US Census Bureau has made a public announcement on their website that the census data will be used for learning about the quality of life of citizens.

Evaluation : Acceptable

If applicable, how is the consent of data subjects obtained?

When filling up the census, citizens were informed that their data will be kept privacy and used with care for learning purposes.

Evaluation : Acceptable

How can data subjects exercise their rights of access and to data portability?

US Citizens can contact for the Chief FOIA Officer at 1-800-432-1494, or by email at pco.policy.office@census.gov. data requests.

Evaluation : Acceptable

How can data subjects exercise their rights to rectification and erasure?

US Citizens can not request data removal from the US Data Census.

Evaluation : Acceptable

How can data subjects exercise their rights to restriction and to object?

Filling the US Data Census is mandatory for US Citizens.

Evaluation : Acceptable

Evaluation : Acceptable

Are the obligations of the processors clearly identified and governed by a contract?

US Census Bureau signed a NDA contract with all Têrepê employees, stating that for a period of 5 years, these could not reveal that they are working with the US Data Census or with the US Census Bureau, neither disclose information about the Census.

Evaluation : Acceptable

In the case of data transfer outside the European Union, are the data adequately protected?

The only processing and storage operations performed by Têrepê are conducted in Portugal. Operations conducted in the US or outside the European Union by the US Census Bureau is not a concern of Têrepê.

Evaluation : Acceptable

Risks

Planned or existing measures

Encryption

At Têrepê offices, Internet access is barred by a secure firewall that makes sure to encrypt any unencrypted connections with TLS 1.1 and 1.2. Outside Têrepê offices, access to the company infrastructure is done by using a VPN using the OpenVPN software. VPN tunnel is protected using IPSec IKE v2.

There is still a risk of disclosure of information by employees when accessing outside the company in unprotected Wi-Fi networks.

Evaluation : Acceptable

Anonymisation

The US Data Census dataset has been anonymized before processing by analytic tools using the ARX software tool.

Evaluation : Improvable

Action plan / corrective actions :

Re-anonymize the dataset to grant that the privacy levels are bigger than the utility levels, but at the same time the analytic tools can provide good results.

Evaluation comment :

Although anonymized, there is still a risk of information disclosure if it revealed to an attacker, as the anonymization had to take into account the needed utility levels for the prediction tool to yield good results.

Logical access control

Data scientists are the only employees who have access to the dataset. A dedicated server to access the dataset was created which can only be accessed using keys distributed to these employees in the form of USB Yubi Keys.

There is still a risk of gaining unauthorized access by internal employees or external attackers, if someone manages to grab/steal the USB Yubi Key.

Evaluation : Improvable

Action plan / corrective actions :

Deploy GPS Security Mechanism to track USB Yubi keys of data scientists

Evaluation comment :

Deploy a GPS security mechanism that allows the track of the USB Yubi keys. This will reduce the likelihood of the lose of proprietary devices.

Paper document security

Product Owners and UI Designers do weekly reports on the tool being created in order to refine its features. For this they use Microsoft Word to create a report, which is then translated in the .PDF format and distributed to board members for presentations and validations. Every document is stored in a secure storage and destroyed after two years.

There is still a risk of disclosure of information by employees if they share these report documents with those that are not on the meetings, either via Social Engineering attacks in E-mails or by talking about it.

Evaluation : Improvable

Action plan / corrective actions :

Deploy a IDS rule to block unrecognized e-mail domains.

Evaluation comment :

Deploy a IDS rule to block unrecognized e-mail domains. This will prevent social-engineering attacks by outsiders.

Network security

CISO implented and deployed a network topology that involves an IDS, a DMZ and firewalls to secure and detect unauthorized accesses.

There is still a risk of disclosure of information if employee install malware or a compromised software.

Evaluation : Improvable

Action plan / corrective actions :

Setup anti install mechanism on employees computers, to prevent install of malware.

Evaluation comment :

Setup anti install mechanism on employees computers, to prevent install of malware. Employees that need to install any software will have to check with the CISO before, so he can validate if it is malware or not.

Risks

Illegitimate access to data

What could be the main impacts on the data subjects if the risk were to occur?

Breach of US Citizens net worth, Public damange, Breach of US Citizens Identity Information, Target of Marketing Companies, Target of Robbers, Political and Social-Economical Power to Russia and China

What are the main threats that could lead to the risk?

Social Engineering, Packet Sniffing, Supply-Chain Attacks, Malware, Proprietary Device Thief, Human Interaction

What are the risk sources?

Employees, Software

Which of the identified planned controls contribute to addressing the risk?

Encrvption. Anonymisation. Logical access control. Paper document security. Network security

How do you estimate the **risk severity**, especially according to potential impacts and planned controls?

Maximum, Breaching data of the citizens of the worlds largest country would be catastrophic, not only because it would put the country citizens identity and status at risk, but also because it would cause enormous fines to the company, Portugal, United States and it would given political advantage to rival countries such as China and Russia.

How do you estimate the **likelihood of the risk**, especially in respect of threats, sources of risk and planned controls?

Important, It is unlikely that the risks may actually happen for two reasons: the company data processing is well secured and the dataset does not direct identifiers, meaning that it would require secret datasets from the US to correlate the data entries.

Evaluation : Improvable

Action plan / corrective actions :

- Deploy GPS security mechanism to reduce likelihood of Proprietary Device Steal;
- Deploy IDS rule to prevent likelihood of Social Engineering via e-mail;
- Block installation of software that is not present on the supported software list;
- Re-anonymize dataset to increase privacy but reduce utility levels, as a measure to reduce the likelihood of correlating the anonymized dataset

Evaluation comment :

Reverted likelihood risk to Limited

Added action plan

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Illegitimate access to data)? **Maximum**

Taking into account the action plan, how do you re-evaluate the **likelihood of this risk** (Illegitimate access to data)? **Limited**

Risks

Unwanted modification of data

What could be the main **impacts on the data subjects** if the risk were to occur?

Public damage

What are the main **threats** that could lead to the risk?

Human Interaction, Proprietary Device Thief, Malware

What are the **risk sources**?

Employees

Which of the identified **controls** contribute to addressing the risk?

Logical access control

How do you estimate the **risk severity**, especially according to potential impacts and planned controls?

Negligible, It will not matter if the employees at Têrepê modify the data, as they are creating a predictive tool that is agnostic of the content of the dataset. The US Census Bureau will use whatever dataset they desire from the census

How do you estimate the **likelihood of the risk**, especially in respect of threats, sources of risk and planned controls?

Negligible, N/A

Evaluation : Acceptable

Evaluation comment :

Revert impact and likelihood risk to Negligible

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Unwanted

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Unwanted modification of data)? **Negligible**

Taking into account the action plan, how do you re-evaluate the **likelihood of this risk** (Unwanted modification of data)? **Negligible**

Risks

Data disappearance

What could be the main impacts on the data subjects if the risk were to occur?

Public damage

What are the main threats that could lead to the risk?

Human Interaction, Malware, Packet Sniffing, Proprietary Device Thief, Social Engineering, Supply-Chain Attacks

What are the risk sources?

Employees, Software

Which of the identified controls contribute to addressing the risk?

Encryption, Logical access control, Network security

How do you estimate the risk severity, especially according to potential impacts and planned controls?

Important, There is no impact if data is lost for the citizens as the US Census Bureau detains the primary source of it. However it would be really bad for the company to lose the dataset, as it would leave a really bad impression for the US Government. It would have a serious impact if that data would be lost in terms of a robbery, but that falls under the *Illegitimate access to data* risk category.

How do you estimate the likelihood of the risk, especially in respect of threats, sources of risk and planned controls?

Negligible, N/A

Evaluation : Improvable

Action plan / corrective actions :

- Improve hardware security mechanisms to protect access to external hard drive in the company facilities.

Evaluation comment :

Added action plan

Taking into account the action plan, how do you re-evaluate the **seriousness of this risk** (Data disappearance)? **Important**

Taking into account the action plan, how do you re-evaluate the **likelihood of this risk** (Data disappearance)? **Negligible**

Risks

Risks overview

Potential impacts

Breach of US Citizens net w...

Public damage

Breach of US Citizens Ident...

Target of Marketing Companies

Target of Robbers

Political and Social-Econom...



Threats

- Social Engineering
- Packet Sniffing
- Supply-Chain Attacks
- Malware
- Proprietary Device Thief
- Human Interaction

Sources

- Employees
- Software

Measures

- Encryption
- Anonymisation
- Logical access control
- Paper document security
- Network security

