International Workshop on Privacy and Security in HealthCare 2014 (PSCare14)

Privacy Threat Modeling for Emerging BiobankClouds

Ali Gholami\textsuperscript{a,*}, Anna-Sara Lind\textsuperscript{b}, Jane Reichel\textsuperscript{b}, Jan-Eric Litton\textsuperscript{c}, Ake Edlund\textsuperscript{a}, Erwin Laure\textsuperscript{a}

\textsuperscript{a}HPCViz and SeRC, KTH Royal Institute of Technology, Stockholm, Sweden
\textsuperscript{b}Faculty of Law and Centre for Research Ethics and Bioethics, Uppsala University, Sweden
\textsuperscript{c}Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Abstract

There is an increased amount of data produced by next generation sequencing (NGS) machines which demand scalable storage and analysis of genomic data. In order to cope with this huge amount of information, many biobanks are interested in cloud computing capabilities such as on-demand elasticity of computing power and storage capacity. There are several security and privacy requirements mandated by personal data protection legislation which hinder biobanks from migrating big data generated by the NGS machines. This paper describes the privacy requirements of platform-as-service BiobankClouds according to the European Data Protection Directive (DPD). It identifies several key privacy threats which leave BiobankClouds vulnerable to an attack. This study benefits health-care application designers in the requirement elicitation cycle when building privacy-preserving BiobankCloud platforms.

1. Introduction

BiobankClouds are gaining popularity due to flexibility and scalability in processing big genomic data. Additionally, BiobankClouds are cost-efficient for biobanks\textsuperscript{1} as commodity hardware ownership is no longer required. The EU FP7 BiobankCloud project [1] aims to build a scalable, reliable, and secure cloud infrastructure for biobanks. This BiobankCloud should support scalable alignment, clustering, aggregation, and compression of genomic data. There are several security concerns that hinder biobanks from capitalizing on the benefits of cloud computing including the risk of unauthorized use of data, loss of control, a multi-tenant environment, and the lack of clear service level agreements (SLAs). Privacy legislation like the EU data protection directive (DPD) [2] and the US health insurance portability and accountability act (HIPAA) [3] mandate stricter requirements for processing sensitive personal health data. These factors impede the migration of existing biobank services to the cloud.

\textsuperscript{*}Corresponding author. Tel.: +46-8-7906356; Fax: +46-8-247784.
E-mail address: gholami@pdc.kth.se

\textsuperscript{1}© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.

Keywords: privacy-preservation; data security; cloud computing; threat modeling; requirement analysis
Privacy threat modeling, an essential stage of secure software engineering, encourages communication of privacy requirements among stakeholders of the biobank when developing privacy-preserving cloud services. There has been considerable development of information security threat modeling frameworks and tools, for example OCTAVE [4] and STRIDE [5]. Unfortunately, the complexity of such frameworks makes them difficult to be applied to a project that demands agile methods with limited resources. It is also worth noting that privacy-preservation is not emphasized in the existing security threat modeling frameworks.

This paper describes the EU DPD privacy requirements and identifies several important privacy threats faced by a BiobankCloud platform as a service (PaaS) model. The privacy threat modeling methodology is based on Cloud Privacy Threat Modeling (CPTM) [6] which adheres to the EU DPD privacy principles. The CPTM methodology provides an agile approach for identifying privacy threats. Furthermore, the CPTM provides guidelines in order to mitigate the effects of these threats for a variety of cloud computing service models within the EU’s jurisdiction. A proof-of-concept of the CPTM methodology provides a threat identification approach for complying with the EU DPD.

This paper contributes the following:

- Classification of the EU DPD privacy requirements for processing the next generation sequencing (NGS) of data in the BiobankCloud.
- Implementation of the CPTM, as a specific cloud privacy threat modeling approach for defining BiobankClouds’ privacy requirements.
- Fine-grain threat identification and ranking of potential adversarial attacks against the privacy of the genomic data.

The structure of this paper is as follows: Section 2 describes related work to highlight the existing research on cloud privacy threat modeling. Section 3, gives a brief overview of the CPTM methodology. Section 4 defines different entities and architecture of a BiobankCloud as well as the key themes of the EU DPD. Section 5 identifies the privacy threats based on the EU DPD requirements. Section 6 presents the conclusions and findings.

2. Background and Related Work

There has been a significant amount of research conducted in the area of threat modeling for various information systems with the goal of identifying a set of generic security threats [7], [8], [9]. Privacy impact assessment (PIA) methodologies have also been developed to assess the impact on privacy for projects, products, policies and services [10].

Extensive guidelines already exist for reducing the security risks of cloud services, but these do not include an outline of privacy threat modeling. The cloud security alliance (CSA) guidelines [11] are not thorough enough to be considered a privacy threat model because they are not specific to privacy-preservation. The European network and information security agency (ENISA) has identified a broad range of security risks and benefits of cloud computing including sensitive data protection [12]. LINDDUN [13] is short for likability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and non-compliance. It proposes a comprehensive generic methodology for privacy requirement elicitation through the mapping of initial data flow diagrams of application scenarios to the corresponding threats. CNIL has proposed a methodology for privacy risk management that information systems requiring DPD may use [14].

Pearson describes the key privacy challenges in cloud computing that arise from a lack of user control, a lack of training and expertise, unauthorized secondary usage, complexity of regulatory compliance, trans-border data flow restrictions, and litigation [15].

1 A biobank is a type of biorepository that stores samples of human biological material for research and clinical services. There are hundreds of well established biobanks world-wide storing, managing and processing bio-specimens. Some are run by public or private institutions, such as hospitals, while in other cases research groups manage their own sample collections. Biobanks also store personal data from the sample donors, such as age, gender, diagnosis, etc.
The above threat modeling methodologies target a wide range of applications, projects, and products. When building cloud computing platforms such methodologies may introduce a significant overhead (both technically and legally) to develop privacy-preserving software systems. One example is a demand for training software designers for several months to learn these methodologies. The CPTM enables us to deliver a privacy threat model for the PaaS BiobankCloud with the focus on the DPD privacy requirements.

3. Cloud Privacy Threat Modeling (CPTM)

The CPTM [6] is a specific privacy-preservation threat modeling methodology for cloud computing environments that process sensitive data within the EU’s jurisdiction. The key differences of the CPTM with the existing threat modeling methodologies are agility through defining relevant DPD requirements, classification of important privacy threats, and providing countermeasures for the identified threats for different cloud computing services (SaaS, PaaS, IaaS) and deployment models (Public, Private, Hybrid, Community).

For the first step in the CPTM approach, the main entities are identified for the developing cloud environment based on the DPD terminology. Secondly, the CPTM describes the privacy requirements (PRs) that must be implemented, e.g., lawfulness, informed consent, purpose binding, data minimization, data accuracy, transparency, data security, and accountability. Finally, the CPTM provides countermeasures for the identified threats against adversarial attacks.

4. BiobankCloud Entities and Privacy Requirements

This section describes the PRs of the BiobankCloud according to the CPTM guidelines. Section 4.1 defines the BiobankCloud entities. Section 4.2 presents private and federated community-based BiobankClouds. Section 4.3 summarizes the DPD requirements that are the basis for threat identification.

4.1. Entities

The main entities (participants) of the DPD are the data subject, controller, and processor. Article 2.a of the DPD defines data subject as an identifiable individual associated with the personal data. In Article 2.d, controller is defined as the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purpose and means of the processing of personal data. The processor acts as the natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller, as defined in Article 2.e. [9].

The DPD participants have been divided into researcher, data provider (DP) and cloud service provider titles (CSP) in the BiobankCloud. The sample donor (research subject, or in the DPD terminology, data subject) cannot be considered an actor in the BiobankCloud since the donor represents the actual genomic data containing the personally identifiable information (PII). The data processed in the cloud is collected from the sample donors via the DP, but the sample donors themselves do not take an active part in the processing.

A. Researcher: there are two categories of researchers: trusted and guest researchers. Researchers in the first group are affiliated with the institutions that hold the genomic data. The trusted researcher acts under the responsibility of the DP, as described in this section. The guest researcher conducts experiments on subjects genomic data. The guest researcher is able to log in to the BiobankCloud and run a workflow on the infrastructure provided by the processor (CSP), but will otherwise have limited capabilities within the Cloud.

B. Data Provider (DP): the DP acts as the controller to permit access to the subjects genomic data through the processor (CSP). The DP may also be responsible for the access granted to the trusted researcher. The guest researcher can only access the BiobankCloud after permission has been granted by the DP.

C. Cloud Service Provider (CSP): the CSP is the entity that performs actual computation and storage of genomic data delegated by the DP. The CSP can be considered the processor, as defined by the DPD, in order to make the BiobankCloud platform available to a set of trusted entities such as researchers and DPs.
4.2. BiobankCloud Architecture

We envisage a BiobankCloud as a PaaS that provides the capability of deploying sequencing applications with their dependencies within an environment called a container. The NGS machines produce large amount of genetic information that will be uploaded to BiobankClouds that are protected by firewalls within CSPs. Such genomic information will be stored in the genomic data storage (GDS) to be accessed by the execution containers (EC) for running the researchers’ experiments, as demonstrated in Fig. 1.

![BiobankCloud Architecture](image)

**Fig. 1.** BiobankCloud Architecture.

There are scenarios where different CSPs form an institutional federation among their existing BiobankClouds. This federation is able to take benefit from each institutions’ capabilities. To achieve this, federated BiobankClouds build a community-based cloud to transfer the genomic data and share the experimental results with their affiliated researchers. Fig. 2, shows a federation between BiobankCloud X and BiobankCloud Y, where they can share genomic data through the organizational boundaries.

![Federation of BiobankClouds to form community-based BiobankCloud](image)

**Fig. 2.** Federation of BiobankClouds to form community-based BiobankCloud.

4.3. Privacy Requirements

In this section, we define the privacy requirements for the BiobankCloud. The DPD fundamental PRs are lawfulness, informed consent, purpose binding, data minimization, data accuracy, transparency, data security, and accountability.

PR1 Lawfulness: sets out the basic premises for the legitimate processing of data, that all processing must be conducted within the regulatory framework of the DPD. Data processing can be allowed on the basis of for

---

2 Paras 18, 23, 28 of the Preamble, Article 6 of the DPD.
example statutory permissions (such as legislation), or with data subject consent, if necessary for the performance of a contract or on the basis of statutory permission such as legislation. In regards to sensitive data such as health data, the DPD holds that the EU Member States must provide further safeguards, for example through involvement of a research ethics committee. Procedures and processes to disclose the sensitive genomic data is governed strictly under the lawfulness requirement. It can also be seen as an umbrella principle, reflecting other requirements and being the point of departure for true data protection.

PR2 Informed Consent:3 informed consent justifies processing of genomic data in the BiobankCloud. The genomic data may have been provided with informed consent through the DP which constitutes the main justification for processing. In cases where data has been collected a long time ago, where the data subject is diseased, or in the case of data on anonymous cell lines, the data may no longer be able to connect to an individual person and thereby fall outside the scope of the DPD. In such case, even non-consented use of data may be permitted. Further there might be some room to re-use previously consented data, if it conforms with the law applicable to the DP and this legislative act indicate the purpose for processing and the purpose is of substantial public interest. This law or the decision must include necessary safeguards so that the interests of the data subjects are effectively protected [16].

PR3 Purpose Binding:4 ensures that personal data processing is performed according to predetermined purposes. The collected genetic data in the BiobankCloud will only be processed according to the purposes covered by the informed consent given by the subject or, if the law applicable to the DP so admits, according to further purposes within the legal framework.

PR4 Data Minimization:5 restricts extra and unnecessary disclosure of information to third parties, such as CSP, to reduce the risk of information leakage that leads to privacy breaches. This requirement of the DPD demands a retention period of the published genetic data to be monitored closely. Storage over time can only be permitted if in accordance with the law applicable to the DP.

PR5 Data Accuracy:6 describes the necessity to keep data accurate and to be updated by the DP. A controller holding personal information shall not use that information without taking steps to ensure with reasonable certainty that the data are accurate and up to date. The obligation to ensure accuracy of data must be seen in the context of the purpose of data processing. In line with the principle of accuracy, data subjects must have the right under national law to obtain from the controller the rectification, erasure or blocking of their data if they think that their processing does not comply with the provision of the directive, in particular because of the inaccurate or incomplete nature of the data. Data accuracy requirement is closely linked with Transparency (PR6), as described in the following.

PR6 Transparency:7 entitles the data subjects to have information about the processing of their data and thereby a means to learn of the processing operation of their data. Transparency thus functions as a prerequisite for the data subjects to monitor that the data is accurate, in accordance to PR5. Transparent data processing is required to be implemented in the BiobankCloud with clear description of technical, physical and organizational measures that CSP has in place to infer if data are processed appropriately.

PR7 Data Security:8 proposes implementing technical measures to provide legitimate access and organizational safeguards. The DP shall ensure that whoever processes the data on his behalf, e.g., the CSP provides adequate levels of security against unlawful data processing. It is stated in Article 17 DPD that Member States shall provide that the controller (data provider) must implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network and against all other unlawful forms of processing. Data security covers the equipment (hardware, software, etc.) but also organizational aspects such as internal rules on how communication with and from the staff of the

---

3 Para 30 of the Preamble, Article 7 of the DPD.
4 Paras 28-31 of the Preamble, Articles 6 and 7 of the DPD.
5 Paras 59-61 of the Preamble, Articles 16-17 of the DPD.
6 Paras 28 and 41 of the Preamble.
7 Paras 38-40 of the Preamble, Articles 10-15 of the DPD.
8 Para 46 of the Preamble, Articles 6, 16-17 of the DPD.
CSP is dealt with, how responsibilities are handled internally etc., and how access to facilities where data is stored is regulated.

**PR8 Accountability** \(^9\): mandates internal, external auditing and control for various assurance reasons. The DP is responsible to ensure compliant of supplied genomic data usage to the CSPs. For instance, to ensure that confidentiality and integrity of data have been preserved by the acting CSP. The CSP shall put in place measures which would under normal circumstances guarantee that data protection rules are adhered to in the context of processing operations; and have documentation ready which proves to DP and to supervisory authorities what measures have been taken to achieve adherence to the data protection rules. This criterion requires the DP to act in a proactive manner, to actively demonstrate compliance and not merely wait for data subjects or supervisory authorities to point out shortcomings.

## 5. BiobankCloud Privacy Threat Model

This section presents the privacy threat model for the BiobankCloud using the CPTM. Section 5.1, describes the platform assets, attackers, and security boundaries in the BiobankCloud. Section 5.2 outlines privacy threat analysis for each class of the PRs.

### 5.1. Assets, attackers and boundaries

The proposed threat model seeks to address privacy and security risks related to genomic data. The BiobankCloud platform containing the genomic data is considered as the asset. Eavesdroppers and malicious users are known as attackers that are able to exploit the possible vulnerabilities in the platform. The security boundaries consist of firewalls that control the incoming/outgoing traffic through a CSP along as well as the physical means to deny access to the computing platform for unauthorized personnel.

### 5.2. Threat analysis

We define the DPD privacy threats according to the CPTM threat classification. The identified threats are indicated as pairs of (a,b). These threats are ranked based on the feedback from the BiobankCloud participants including hospitals, biobanks, technology providers, and researchers during requirement elicitation phase. The first parameter of the pair indicates probability occurrence of a threat while the second parameter indicates the importance factor of exploiting the vulnerability associated with that threat. The proposed values are low (L), moderate (M) and high (H). For instance, T1,j(M,H) means threat number Ti,j has a moderate (M) occurrence and it significantly (H) affects the PRs.

Such ranking facilities communication of threats and their importance among the project members and also prioritization of threats. It is important to notice that the rankings are developed for the BiobankCloud project and there is not a sound mathematical proof for its correctness. This model needs to be updated individually for each developing platform and actually can be different for other environments/operators. For instance, some configurations may also require ranking of exposure or compliance with additional privacy legislation such as HIPPA act.

**T1** PR1 (Lawfulness) PR1 can be considered an umbrella principle. In this context, any violation the PRs may amount to a lawless threat. A non-exhaustive list of the major threats of PR1 are identified as follows:

- **T1.1(L,H):** Lack of relevant information on legal rights and duties to allow data subject and other interested parties to use effective means for accessing the data.
- **T1.2(L,M):** Amendments of legal requirements or unawareness of new rules. Incorrect interpretation or application of legal concepts leading to unlawful processing of (complex) data; for instance a DP categorizing genomic data as non-sensitive.

\(^9\) Paras 55-64 of the Preamble, Articles 22-24 of the DPD.
• T1.3(L,L): Unawareness of responsibilities or legal requirements due to unclear contracts/ terms of service (ToS).
• T1.4(M,H): Lack of agreement from all entities regarding the processing the genomic data.
• T1.5(L,M): If data are not obtained in accordance with the law, or without approval (informed consent or ethics board).

T2 PR2 (Informed Consent)

• T2.1(H,H): Excessive ToS containing too much specific information that are not clear enough, e.g., legal terms that are not easily understandable for a layman. This could result in data subject not being adequately informed before giving their consent.
• T2.2(H,M): Lack of possibility to give consent dynamically to a specific subset of genomic data.

T3 PR3 (Purpose Binding)

• T3.1(L,H): Researcher does not use the data according to the initial purpose or the DP does not use the data according to the purposes.
• T3.2(M,H): Researcher who has access to multiple data studies, makes cross-link analysis to the genomic data that is not consented and, hence illegal, according to the PR1 and PR2.

T4 PR4 (Data Minimization)

• T4.1(L,L): The requested sensitive data to be used by CSP or guest researcher is not certain or well defined in advance. For instance, additional sensitive attributes that are not necessary to be included in the published microdata to be used by a researcher.
• T4.2(M,M): If the DP does not define the retention period of the sensitive (genomic) data, there is a threat of accumulating more and more sensitive data over time, that can be used for inference and linking attacks.

T5 PR5 (Data Accuracy)

• T5.1(L,M): If the DP cannot or will not update the data when having wrong information has been found to be incorrect.
• T5.2(L,L): DP uploads the data to the CSP but the data source validity is not affirmed.

T6 PR6 (Transparency)

• T6.1(H,M): Lack of communication and information between entities. The threat is even more sever when the lack of information or openness is harmful for weaker parties, the sample donor/data subject.
• T6.2(L,M): Researchers or DP cannot get access to modify or erase data, due to unclear data processing procedures.

T7 PR7 (Data Security)

• T7.1(M,H): Theft of authentication credentials by an adversary through phishing attacks or network eavesdropping, brute force attacks to guess authentication credentials or identities of users.
• T7.2(L,M): Repudiation of access to the genomic data by the researchers.
• T7.3(M,H): Wide access to data by a large group of people.
• T7.4(M,H): Elevation of privileges by an attacker to change the access rights to higher privileges.
• T7.5(L,H): Insecure flow of information between CSPs in a community-based BiobankCloud.
• T7.6(M,H): Theft of genomic data at rest or during runtime by a malicious insider.
• T7.7(L,M): Unavailability of data due to denial of service (DoS) attacks.
• T7.8(M,M): Session reply through message theft by eavesdropping to steal a session.
• T7.9(L,M): Theft of private keys by an attacker to decrypt the sensitive genomic data.
• T7.10(L,M): Inference attacks through guessing or background knowledge to the minimized data, required by PR4.
• T7.11(M,H): Storing passwords, credentials, database connections, keys in plain text or within the source code.

T8 PR8 (Accountability)

• T8.1(M,M): Lack of mechanism for secure auditing to provide evidence of confidentiality and integrity.
• T8.2(L,M): Lack of awareness or routines to implement data protection rules.
• T8.3(L,M): Logging information or audit trails contain sensitive information about the subjects.
• T8.4(H,M): Excessive information in the audit logs to make audit and inspection about usage of the genomic data by an auditor.

6. Conclusions

The Cloud Privacy Threat Modeling (CPTM) defines an agile platform development approach while taking into consideration the DPD and potential privacy threats. A privacy threat model for emerging cloud services is defined for both private and community-based cloud deployment models for biobanks. The proposed threat model facilitates communication of privacy requirements for the BiobankCloud.

To this aim, the CPTM is used for privacy threat modeling for cloud environments. Each privacy requirement category is identified and ranked based on feedback from a variety of biobank experts. The results of this study are to be used in the design of the BiobankCloud infrastructure in the BiobankCloud project [1].

Acknowledgements

This work partially funded by the EU FP7 project Scalable, Secure Storage and Analysis of Biobank Data under Grant Agreement no. 317871.

References