Privacy Requirements Report

Francesco Paolo Liuzzi
Team A - Topic 7

October 2024

Contents

1 Introduction

This report outlines the basic privacy requirements to be met by the system and is based on current regulations, including the General Data Protection Regulation (GDPR)[1], as well as best privacy-oriented software engineering practices (Privacy-Oriented Software Design, POSD) [2].

For each requirement, a reference to relevant regulations or literature is provided and, where applicable, also the main methodologies and tools (software, libraries, etc.) recommended for their effective implementation. It should be noted that these tools are recommended, but not mandatory. Developers are therefore free to use the tools they consider most appropriate and of equivalent effectiveness, taking into account the various constraints, including the cost. However, this does not exclude the obligation to comply with the requirements outlined in this document, which aim to ensure the implementation of a system that complies with the main privacy principles established by the current legislation.

Compliance with these requirements will subsequently be verified and if flaws are found, they will be communicated to the responsible team.

2 GDPR

This section outlines the application of the General Data Protection Regulation (GDPR) principles to the project. The GDPR, a comprehensive data privacy law implemented by the European Union, is designed to protect individuals' data and ensure that organizations process this data transparently, fairly, and securely. The regulation requires organizations to handle personal data with strict respect for privacy rights, including the right to access, correct, and erase data.

Under GDPR, personal data includes any information that can directly or indirectly identify an individual, such as names, email addresses, IP addresses, and even data generated by users' activities that could link back to them. Even if not all the data of the project can be considered PII, we redact the following report proactively so that the system will be compliant also in case of future changes in the data type which it will deal with.

We would like to clarify that, following the overall analysis conducted for the implementation of this project, all data provided and produced by the SMO and FFM, as well as any other logged access to the platform, have been identified as personal data. As noted above, this report is proactive, in line with best practices, in dealing with any non-drastic changes (i.e. not involving a full overhaul of the requirements) to the behavior of the platform.

In the following sections, we apply the seven core data protection principles outlined by the GDPR to guide data processing practices in the project. These principles—Lawfulness, Fairness, and Transparency; Purpose Limitation; Data Minimization; Accuracy; Storage Limitation; Integrity and Confidentiality; and Accountability—provide a framework to ensure that all data collected and processed are managed responsibly and with respect for users' privacy rights. This approach will ensure that the project not only complies with GDPR but also upholds best practices for data privacy and protection software design.

2.1 Lawfulness, fairness, and transparency

Every data processing activity within the system must have a legal basis. Common legal bases include user consent, the performance of a contract, legitimate interests, compliance with legal obligations, or protection of vital interests (GDPR Art. 6). This involves making it clear to users what data is collected, why it is collected, and how it will be processed.

Example. *If you provide a way to create user accounts, specify why you request some data especially if the field is mandatory.*

User consent is mandatory if data processing extends beyond originally stated functions, such as using data for marketing analytics or employee performance tracking.

Users must be informed about how their data is collected, processed, and used, especially if automated processing or profiling is involved (GDPR Articles 12-14).

Example. *If you start using personal data for marketing purposes, you should inform the users involved (e.g., by sending an informative email).*

2.1.1 How to

1. **Legal Basis for Processing**
    - **Purpose Documentation:** publish a detailed privacy policy that describes the types of data collected, purposes of processing, data retention, and data sharing practices.
    - **API Consent Management:** APIs should only allow data retrieval or processing if the user has provided explicit consent. Use a consent management platform integrated with API endpoints to control data processing based on user consent dynamically.
2. **User Notification and Disclosures**
    - Display privacy policies prominently at data collection points, explaining the purpose, retention period, and any third-party sharing practices.
    - **Data Transparency Dashboard:** build a user-facing dashboard in the app where users can view which data is being processed, the purpose, and by which AI models. Track access logs to provide accountability and detect misuse.
3. **Data Queries Transparency**

  - **Audit Trails:** enable audit logging in databases to track all read/write operations. Include logs for each query executed by APIs to ensure transparency in data access.
  - **Field-Level Permissions**: apply field-level encryption or access control to sensitive database fields, ensuring API and AI models access only the authorized data fields relevant to their processing task.

2.1.2 Tools

- **Privacy Policy Generators and Consent Management**
  - **OneTrust**: manages consent across APIs, enables customizable privacy policies, and tracks user consent.
  - **TrustArc**: provides consent management for compliance with GDPR, supporting multi-language and region-specific consent flows.
- **Data Access and Transparency**
  - **Okta[3]**: allows for secure, user-specific access and consent validation for data usage in APIs.
  - **Privacy Patterns**: use patterns from PrivacyPatterns.org to guide UI transparency practices for disclosing data usage.

2.2 Purpose Limitation

Data must only be used for specified, legitimate, and explicit purposes. Any deviation from the initial purpose must be consented to by the user or justified by a valid legal basis (GDPR Article 5(1)(b)).

2.2.1 How to

1. **Data Processing Register**
   - **Document Processing Purposes:** include details for each data type, the associated legal basis, and justification of purpose. Make this register available to data protection authorities if requested.
   - **Regular Updates:** periodically review and update user consents if additional data processing or new functionalities are introduced.
2. **Database Segmentation**
   - **Data Segmentation by Purpose:** structure databases with separate tables, schemas, or views based on purpose (e.g., analytics, user profiling) to prevent unauthorized cross-purpose processing.
   - **Tagged APIs Data Fields:** tag API-accessible data fields by purpose, ensuring that data retrieved through APIs is limited to its intended purpose. Use a policy engine to enforce API requests that align with the documented purpose.

3. **AI Purpose-Limited Processing**

    - **Purpose-Specific AI Models:** train AI models on datasets specifically prepared for the intended use, preventing cross-purpose data processing. Use explainability tools to validate that AI outcomes adhere to the original purpose.
    - **Model Drift Monitoring:** implements drift detection to ensure that AI models remain within the intended processing scope, alerting administrators if models begin processing outside the defined parameters.

4. **Access Control by Purpose in APIs**

    - **Role-Based Purpose Access:** restrict API access based on purpose and user role.
    - **Dynamic Purpose Controls in APIs:** configure APIs with dynamic checks to reject requests that don't align with the purpose, logging any purpose violations for auditing.

2.2.2 Tools

- **Policy Enforcement and API Gateways**
    - **Open Policy Agent (OPA)**[4]: enforces purpose-based access policies by validating each data request against documented purposes.
    - **Kong API Gateway**[5]: manages API endpoints and enforces purpose-specific data access controls through plugins.
- **Database Segmentation and Purpose Tagging**
    - **PostgreSQL Schemas and Views**: segment data by purpose at the database level using schemas or views to keep datasets separate and enforce purpose limitation.
    - **Snowflake Data Governance**[6]: manages data access with built-in tagging and purpose-specific controls for data warehouses.

2.3 Data Minimization

Only the data necessary for each purpose must be collected and processed, preventing the excessive gathering of information. (GDPR Art. 5(1)(c))

Example. *When setting up the alert system for production costs, limit data collection to relevant cost indicators without unnecessary personal identifiers or unrelated metrics.*

2.3.1 How to

1. **Database Field Minimization**

   - **Field-Level Controls and Constraints:** in databases, enforce field constraints that limit the type and amount of data stored per purpose (e.g., limit date fields to month/year when the full date isn't necessary).
   - **Dynamic Data Views:** use views or dynamic queries in databases to selectively expose only the minimum data needed for specific API calls.

2. **API Data Filtering**

   - **Granular API Responses:** design APIs to return only the requested data points.
   - **Field-Level Data Minimization**: use request validation libraries in APIs to filter out any extraneous data, enforcing the minimal data collection principle.

3. **AI Data Minimization**

   - **Pseudonymization**: implement pseudonymization for datasets used in AI training, masking or removing identifiers.
   - **Conditional Feature Selection:** use feature selection methods to limit the model to the most essential features, reducing the amount of personal data processed.

2.3.2 Tools

- **Privacy-Enhancing Technologies (PETs)**

  - **privacyIDEA**[7]: an open-source solution for enforcing minimal data collection with role-based access control, beneficial for managing multi-factor authentication.
  - **ARX Data Anonymization Tool**[8]: anonymizes sensitive data, allowing non-identifiable datasets for analytics while minimizing exposure.

- **Data Validation and Minimization in APIs**

  - **GraphQL**: allows for selective data fetching, ensuring only necessary data fields are retrieved by client applications.
  - **Apache Commons Validator**[9]: validates input fields, allowing only essential data to be entered or processed in forms and APIs.

- **Conditional Feature Selection in AI**

  - **Scikit-Learn's Feature Selection**: allows selective use of features in model training, limiting data exposure by only processing essential fields.

2.4 Accuracy

Personal data must be accurate, and every reasonable step should be taken to ensure that inaccurate data is corrected. (GDPR Art. 5(1)(d)).

2.4.1 How to

1. **Database Validation Constraints**
   - **Integrity Constraints**: apply integrity constraints (e.g., unique, not null) on critical fields to maintain data accuracy. Use database triggers to enforce validations when data is updated or inserted.
   - **Data Quality Monitoring**: set up automated data validation and accuracy checks in databases using data quality platforms to verify accuracy regularly.
2. **API Data Validation**
   - **Real-Time Data Validation**: use validation libraries (e.g., Apache Commons Validator or JSON Schema for API responses) to enforce data format and accuracy requirements at the input level.
   - **Error Handling and Correction Mechanism**: implement error correction in API responses, allowing users to be notified of data inconsistencies and enabling correction processes when needed.
3. **AI Model Accuracy Checks**
   - **Recalibration**: routinely retrain AI models on updated data to maintain accuracy. Track model performance and identify drift.
   - **User Feedback Loops**: integrate a feedback loop where users can report inaccuracies, which then triggers a review or retraining of AI models if needed.

2.4.2 Tools

- **Data Quality Tools**
  - **Talend Data Quality**[10]: automates data validation and quality checks for databases, reducing inaccuracies in large datasets.
  - **Informatica Data Quality**[11]: integrates with databases to ensure consistent data validation and quality rules are applied across the system.
- **Static Code Analysis (SCA) Tools**

  - **Fortify SCA**[12]: detects code-level vulnerabilities and logical flaws that may lead to data inaccuracies.
  - **SonarQube**[13]: analyzes code to ensure that data handling is consistent with accuracy requirements, identifying potential bugs that could affect data quality.

- **Feedback Loops for AI**
  - **MLflow**: tracks machine learning model accuracy over time and integrates user feedback to maintain model accuracy.
  - **TensorFlow Model Analysis (TFMA)**[14]: enables model retraining and re-evaluation, ensuring AI models maintain accuracy based on updated data.

2.5 Storage limitations

Data must be retained only as long as necessary and should be securely deleted or anonymized when no longer needed. (GDPR Art. 5(1)(e))

2.5.1 How to

1. **Database Retention Policies and Automatic Deletion**
   - **Retention Schedules**: set database retention policies (e.g., MySQL EVENT scheduler, PostgreSQL cron jobs) to delete or archive data after a predefined period.
   - **Data Anonymization for Archiving**: use anonymization libraries like ARX to anonymize data required for long-term retention, maintaining analytical use without personal identifiers.
2. **API and Data Access Expiry**: ensure API access tokens have expiration policies, so data can't be indefinitely accessed. Refresh tokens should be limited based on necessity.
3. **Data Expiry Tags**: implement expiration tags in API responses to notify users when data will be deleted, ensuring transparent retention practices.
4. **AI Model Data Lifespan Management**
   - **Time-Limited Model Training Data**: limit the time for which AI models retain specific training datasets, retraining them periodically with fresh, minimized data to avoid retaining outdated information.

2.5.2 Tools

- **Automated Data Retention and Deletion**
  - **Apache NiFi**[15]: Automates data flow management, including data retention and deletion schedules based on regulatory requirements.

  - **AWS S3 Lifecycle Policies**: Defines retention periods and automated deletions for stored data in Amazon S3, especially useful for archival storage with GDPR-aligned retention practices.

- **Anonymization and Data Archival**

  - **DataGrail**[16]: Helps automate data retention and deletion workflows, enforcing GDPR-compliant data lifecycle management.
  - **Google BigQuery**: Offers data retention policies and automated expiration settings for datasets that can be configured based on GDPR storage limits.

2.6 Integrity and confidentiality

Ensure data security to prevent unauthorized access, disclosure, or loss (GDPR Art. 5(1)(f)).

2.6.1 How to

1. **Database Security**

   - **Encryption in Transit and at Rest**: encrypt sensitive data in databases using AES-256 for data at rest and enforce TLS 1.3 for data in transit.
   - **Database Access Controls**: apply role-based access controls (RBAC) with strong authentication measures, such as multi-factor authentication (MFA), to limit database access to authorized personnel.

2. **API Security Control**

   - **API Auth**: use OAuth 2.0 and JSON Web Tokens (JWT) for secure API authentication, ensuring only verified requests can access data.
   - **Rate Limiting and Throttling**: prevent unauthorized access by setting rate limits on APIs using API gateways to mitigate potential data scraping or attacks.

3. **AI Data Security and Confidentiality**

   - **Federated Learning**: where feasible, apply federated learning to keep data within user devices while still training models, reducing the risk of sensitive data exposure.
   - **Encrypted Model Storage**: encrypt models when stored and in transit, ensuring unauthorized parties cannot access model data.

2.6.2 Tools

- **Encryption and Key Management**
  - **HashiCorp Vault**[17]: manages and protects encryption keys, allowing for consistent data encryption at rest and in transit.
  - **OpenSSL**: implements TLS encryption for data in transit, securing API communication and preventing interception.
- **Access Control and Authentication**
  - **Auth0**: provides role-based access controls and multi-factor authentication, ensuring only authorized personnel can access sensitive data.
  - **Keycloak**[18]: open-source identity and access management system that integrates with APIs and databases to enforce secure access.

2.7 Accountability

Demonstrate compliance with GDPR through documented processes, regular audits, and transparent reporting mechanisms (GDPR Art. 5(2)).

2.7.1 How to

1. **API and DB Audit Trails:**
   - enables comprehensive audit logging in databases to record access and modification and ensure APIs log all requests and responses for compliance.
   - **Centralized Audit Management:** use a centralized logging tool to monitor and analyze logs across databases and APIs for suspicious activities and maintain records.
2. **AI Accountability and Transparency**
   - **Model Documentation:** document AI model data sources, processing methods, and outcomes, with clear logs to show adherence to GDPR principles.
   - **Model Impact Assessment:** Conduct regular impact assessments for AI processing to evaluate potential risks and adjust models for privacy compliance. Use tools to audit fairness and accuracy.
3. **Documentation and Reporting Software**
   - **Compliance Reporting Software:** leverage compliance tools to create automated GDPR compliance reports, detailing adherence to data protection principles across the system.

2.7.2 Tools

- **Audit Logging and Monitoring**
  - **Splunk**[19]: collects, monitors, and analyzes audit logs, providing a clear accountability trail for all data access and processing actions.
  - **ELK Stack (Elasticsearch, Logstash, Kibana)**[20]: aggregates logs and monitors system activity, enabling detailed traceability for GDPR compliance audits.
- **Compliance Reporting and Documentation**
  - **OneTrust**[21]: automates GDPR compliance tracking, providing audit reports and data mapping to support accountability.
  - **Netwrix Auditor**[22]: monitors and reports on access, changes, and compliance in databases, file servers, and cloud environments, creating an audit trail for data processing activities.
- **PKB Documentation**
  - **Confluence**: for maintaining the PKB, Confluence can store and organize privacy-related documentation and decisions, ensuring that GDPR compliance efforts are traceable.
  - **JIRA**: tracks privacy tasks, design decisions, and audit requirements, creating an organized accountability framework for ongoing compliance activities.

2.8 User rights

Be aware that the user may exercise the following rights at any time, if applicable.

1. **Right to Access (GDPR Art. 15)** Individuals have the right to access their personal data and obtain information on how it is being processed.

   Key Points:

   - Individuals can request a copy of their personal data.
   - Organizations must confirm whether personal data is being processed and provide access to that data.
   - Information provided should include the purposes of processing, categories of data involved, and recipients of the data.
   - Response should generally be within one month of the request unless complex situations arise (which may allow an extension).

2. **Right to Rectification (GDPR Art.16)** Individuals have the right to have inaccurate personal data corrected and incomplete data completed.

- This includes both factual inaccuracies (like a misspelled name) and outdated information.
- Organizations should respond to requests without undue delay.
- Organization should also inform any third parties with whom the data has been shared about these corrections, if possible.

3. **Right to Erasure (GDPR Art.17)** Individuals have the right to request the deletion of personal data in certain situations. It applies if:

   - Data is no longer necessary for the original processing purpose.
   - Consent is withdrawn and there's no other legal ground for processing.
   - Data is unlawfully processed.
   - Data needs to be erased to comply with a legal obligation.

   This right is not absolute. For example, to comply with legal obligations one can retain necessary data for law enforcement purposes.

4. **Right to Object (GDPR Art. 21)** Individuals have the right to object to the processing of their personal data in certain cases, including direct marketing.

   Note that for direct marketing, the right to object is absolute. Upon objection, processing for these purposes must cease immediately.

All of these rights are supported by GDPR Art. 12 which mandates that information related to these rights be provided in a clear, transparent, and easily accessible way.

2.9 AI-based Service Providers Note

If you intend to use LLMs or other AI-related services on the cloud, be sure that the company is in the Data Privacy Framework list. If not, try to use their services from an alternative provider which is on the list, if possible.

Example. *OpenAI is not on the DPF list. Regardless of whether API endpoints are in Europe or not it is good to pay attention to what data you provide to the service. For greater confidentiality and compliance, the same models can be used through Microsoft's Azure platform, which adheres to the DPF.*

3 Other EU Regulations

The following section delineates specific requirements derived from other key EU legislative frameworks for both personal and non-personal data.

3.1 Data Governance Act (DGA)

The DGA[23] addresses both personal and non-personal data. It aims to facilitate data sharing across sectors while ensuring compliance with existing data protection laws. When personal data is involved, the General Data Protection Regulation (GDPR) applies.

- **Data Intermediation Services (Article 10)** Implement secure and standardized mechanisms for data sharing between entities. Entities can be:
  - **Data Holders**: Individuals or organizations that possess data and are authorized to share it.
  - **Data Users**: Entities seeking to access and utilize data for purposes such as analysis, research, or service development.
  - **Data Subjects**: Individuals whose personal data is being processed.

3.2 Data Act

The Data Act[24] focuses on non-personal data, promoting access and sharing to foster innovation and competition. It complements the DGA by providing a framework for non-personal data utilization.

- **Data Sharing Obligations (Art.4)** Users should have the ability to access and share data they generate, promoting transparency and user empowerment.
- **Third-Party Data Access (Art.5)** Develop secure authentication and authorization mechanisms within your APIs to manage third-party access, ensuring data integrity and security.
- **Dispute Resolution Mechanism (Art. 10)** Establish clear procedures for resolving disputes over data access and usage rights to address conflicts and maintain trust among parties.

3.3 Free Flow of Non-Personal Data

This regulation [25] exclusively pertains to non-personal data, prohibiting unjustified data localization requirements and ensuring the free movement of non-personal data within the EU.

- **Data Localization Prohibition(Art.4)** Avoid unjustified data localization requirements, allowing data to be stored and processed anywhere within the EU.
- **Data Portability Facilitation (Art.6)** Implement measures to facilitate the portability of non-personal data between different IT systems and cloud services.

- **Regulatory Data Access (Art.5)** Ensure that competent authorities have access to data for regulatory control purposes, even if the data is stored in another member state.

References

[1] European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Official Journal of the European Union, L 119, 4 May 2016. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[2] Maria Teresa Baldassarre et al. "Privacy Oriented Software Development". In: *Quality of Information and Communications Technology. QUATIC 2019.* Ed. by Mario Piattini et al. Vol. 1010. Communications in Computer and Information Science. Springer, Cham, 2019, pp. 18–30. DOI: 10.1007/978-3-030-29238-6_2. URL: https://doi.org/10.1007/978-3-030-29238-6_2.

[3] Okta, Inc. *Okta - Identity and Access Management Solutions.* URL: https://www.okta.com.

[4] Open Policy Agent. *Open Policy Agent: Policy-based Control for Cloud-Native Environments.* URL: https://www.openpolicyagent.org.

[5] Kong, Inc. *Kong Gateway - The World's Most Popular Open Source API Gateway.* URL: https://konghq.com/gateway/.

[6] Snowflake Inc. *Snowflake Data Governance - Secure and Govern Your Data in the Cloud.* URL: https://www.snowflake.com.

[7] NetKnights GmbH. *PrivacyIDEA - Open Source Two-Factor Authentication System.* URL: https://www.privacyidea.org.

[8] ARX GmbH. *ARX Data Anonymization Tool - Open Source Software for Data Anonymization.* URL: https://arx.deidentifier.org.

[9] The Apache Software Foundation. *Apache Commons Validator - Data Validation for Java Applications.* URL: https://commons.apache.org/proper/commons-validator/.

[10] Talend S.A. *Talend - Cloud Data Integration and Data Integrity.* URL: https://www.talend.com.

[11] Informatica LLC. *Informatica - Intelligent Data Management Cloud.* URL: https://www.informatica.com.

[12] Micro Focus. *Fortify Software Composition Analysis - Identify and Manage Open Source Vulnerabilities.* URL: https://www.microfocus.com/en-us/cyberres/application-security/software-composition-analysis.

[13] SonarSource SA. *SonarQube - Continuous Code Quality and Security.* URL: https://www.sonarqube.org.

[14] TensorFlow Team. *TensorFlow Model Analysis - Evaluate and Analyze TensorFlow Models.* URL: https://www.tensorflow.org/tfx/model_analysis.

[15] The Apache Software Foundation. *Apache NiFi - Data Flow Automation.* URL: https://nifi.apache.org.

[16] DataGrail, Inc. *DataGrail - Privacy Management Software for Data Privacy Compliance.* URL: https://www.datagrail.io.

[17] HashiCorp, Inc. *HashiCorp Vault - Secure Access to Secrets and Sensitive Data.* URL: https://www.hashicorp.com/products/vault.

[18] Red Hat, Inc. *Keycloak - Open Source Identity and Access Management.* URL: https://www.keycloak.org.

[19] Splunk Inc. *Splunk - Data Platform for Observability, Security, and IT Operations.* URL: https://www.splunk.com.

[20] Elastic NV. *ELK Stack - Elasticsearch, Logstash, and Kibana for Search and Analytics.* URL: https://www.elastic.co/what-is/elk-stack.

[21] OneTrust. *OneTrust - Privacy, Security, and Data Governance Software.* URL: https://www.onetrust.com.

[22] Netwrix Corporation. *Netwrix - Data Security Platform for IT Teams.* URL: https://www.netwrix.com.

[23] European Parliament and Council of the European Union. *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance (Data Governance Act).* 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868.

[24] European Parliament and Council of the European Union. *Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act).* 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0067.

[25] European Parliament and Council of the European Union. *Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a Framework for the Free Flow of Non-Personal Data in the European Union.* 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1807.