Report on security requirements for FAIR implementation and data reusability in health research
DISCLAIMER: Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

COPYRIGHT MESSAGE: © FAIR4Health Consortium, 2019

This report contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.
### Document Information

<table>
<thead>
<tr>
<th>GRANT AGREEMENT NUMBER 824666</th>
<th>ACRONYM: FAIR4Health</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full title</strong></td>
<td>Improving Health Research in EU through FAIR Data</td>
</tr>
<tr>
<td><strong>Horizon 2020 Call</strong></td>
<td>SwafS-04-2018: Encouraging the re-use of research data generated by publically funded research projects</td>
</tr>
<tr>
<td><strong>Type of action</strong></td>
<td>Research and Innovation Action</td>
</tr>
<tr>
<td><strong>Start Date</strong></td>
<td>1st December 2018</td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td>36 months</td>
</tr>
<tr>
<td><strong>Website</strong></td>
<td><a href="http://www.fair4health.eu">www.fair4health.eu</a></td>
</tr>
<tr>
<td><strong>Project Officer</strong></td>
<td>Raluca Iagher</td>
</tr>
<tr>
<td><strong>Project Coordinator</strong></td>
<td>Carlos Luis Parra Calderón, Andalusian Health Service</td>
</tr>
<tr>
<td><strong>Report</strong></td>
<td>Report on security requirements for FAIR implementation and data reusability in health research</td>
</tr>
<tr>
<td><strong>Related task</strong></td>
<td>T2.3 Security requirements for FAIR implementation and data reusability in health research</td>
</tr>
<tr>
<td><strong>Release date</strong></td>
<td>May 2019</td>
</tr>
<tr>
<td><strong>Dissemination Level</strong></td>
<td>Public</td>
</tr>
<tr>
<td><strong>Responsible Author</strong></td>
<td>Francisco J. Núñez Benjumea</td>
</tr>
<tr>
<td><strong>e-mail</strong></td>
<td><a href="mailto:Francisco.nunez.exts@juntadeandalucia.es">Francisco.nunez.exts@juntadeandalucia.es</a></td>
</tr>
<tr>
<td><strong>Contributors</strong></td>
<td>Vasiliki Foufi, Marco Schweitzer, Lukas Huber, Anil Sinaci, Alexander Hörbst</td>
</tr>
<tr>
<td><strong>Keywords</strong></td>
<td>Cybersecurity, FAIR data, health data reusability</td>
</tr>
</tbody>
</table>
Table of Contents

1. Introduction .......................................................................................................................... 2
  1.1. Common Types of Cybersecurity ......................................................................................... 2
  1.2. Cybersecurity measures ......................................................................................................... 3

2. FAIR4Health scenarios related to cybersecurity ....................................................................... 4
  2.1. Scenario 1: Local access to raw data .................................................................................. 4
  2.2. Scenario 2: Local access to FAIR data ................................................................................. 5
  2.3. Scenario 3: PPDDM access to FAIR data ............................................................................. 5
  2.4. Scenario 4: P2P access to FAIR data .................................................................................. 6
  2.5. Scenario 5: Access to eHealth services .............................................................................. 6

3. FAIR4Health Security requirements ....................................................................................... 7
  3.1. General Security requirements ........................................................................................... 8
  3.2. Specific security requirements ............................................................................................ 10

4. Technical recommendations related to scenarios identified .................................................. 11
  4.1. General Concerns ................................................................................................................ 11
  4.2. Identity Management Module ............................................................................................. 12
  4.3. Authentication & Authorization Module (Auth Module) ..................................................... 12
  4.4. Audit Module ....................................................................................................................... 13
  4.5. PPDDM Security Module ..................................................................................................... 14
  4.6. End-to-end Security Module ............................................................................................... 14
  4.7. Utilization of the Cybersecurity modules within the scenarios ........................................... 14

5. Conclusions .......................................................................................................................... 17
  5.1. Technical and functional recommendations ......................................................................... 17

References .................................................................................................................................. 18
1. Introduction

"As modern technology becomes indispensable in health care, the vulnerabilities to cyber-threats continue to increase, compromising the health information and safety of millions of people. This threat can happen in several ways: data can be stolen; data might be deleted or corrupted in a way that is not obvious until years later; and medical devices can be hacked, causing direct harm to patients." [1]. One of the most serious attacks has been the WannaCry attack that encrypted data and files on 230 000 computers in 150 countries and impaired the functionality of the National Health Service (NHS) in England among other organisations [1,2]. Also, "the USA has had some of the most highly publicized cyberattacks on health care; in 2015, criminals stole 80 million records from Anthem, a US health insurance company." [1,3].

Cybersecurity (also referred to as information security) refers to the practice of ensuring the integrity, confidentiality, and availability (ICA) of information. Cybersecurity is comprised of an evolving set of tools, risk management approaches, technologies, training, and best practices designed to protect networks, devices, programs, and data from attacks or unauthorized access [4].

In the last 3 years (2016-2019), two systematic reviews on cybersecurity in healthcare have been published [5,6]. The first one was conducted through the CINAHL, Academic Search Complete, PubMed, and ScienceDirect databases to gather literature relative to cyber threats in healthcare and resulted that the most prevalent cyber-criminal activity in healthcare is identity theft through data breach. Other concepts identified in this research are internal threats, external threats, cyber-squatting, and cyberterrorism [5].

The second one was performed through the CINAHL and PubMed (MEDLINE) and the Nursing and Allied Health Source via ProQuest databases. The limitation of the study was that it focused geographically on the American healthcare system and on only English articles and articles relevant to the U.S. health system. This study reached to the conclusion that "the healthcare industry is a prime target for medical information theft as it lags behind other leading industries in securing vital data" and that "it is imperative that time and funding is invested in maintaining and ensuring the protection of healthcare technology and the confidentiality of patient information from unauthorized access" [6].

1.1. Common Types of Cybersecurity

The most common types of cybersecurity are the following [4]:

a) **Network Security** protects network traffic by controlling incoming and outgoing connections to prevent threats from entering or spreading on the network.

b) **Data Loss Prevention (DLP)** protects data by focusing on the location, classification and monitoring of information at rest, in use and in motion.

c) **Cloud Security** provides protection for data used in cloud-based services and applications.
d) Intrusion Detection Systems (IDS) or Intrusion Prevention Systems (IPS) work to identify potentially hostile cyber activity.

e) Identity and Access Management (IAM) use authentication services to limit and track employee access to protect internal systems from malicious entities.

f) Encryption is the process of encoding data to render it unintelligible and is often used during data transfer to prevent theft in transit.

g) Antivirus/anti-malware solutions scan computer systems for known threats. Modern solutions are even able to detect previously unknown threats based on their behavior.

1.2. Cybersecurity measures

To combat cybersecurity threats in health care, Estonia introduced blockchain technology to securely manage electronic patient records creating a time-stamped record of anyone coming in contact with it and adding or omitting information [1]. The most recent development has been the launch of the US National Institute of Standards and Technology cybersecurity framework in 2018, which represents a collaboration between the US government and private entities to improve cyber infrastructures throughout the country [1].

The HIPAA Security Rule – Focuses on securing the creation, use, receipt, and maintenance of electronic personal health information by HIPAA-covered organizations. The Security Rule sets guidelines and standards for administrative, physical, and technical handling of personal health information. HIPAA implemented physical and technical safeguards to ensure sensitive information is protected from cybercriminals. Physical safeguards include workstation use and security, device and media controls, and facility access controls. Technical safeguards include a unique user identification number, emergency access procedure, automatic logoff, encryption, and decryption.

Having strong passwords is an important step for all organizations, according to the Office for Civil Rights (OCR). This means having a password that is 10 characters or longer, and involves a combination of upper and lowercase letters, numbers, and special characters. Entities should also practice good password hygiene, and not share their passwords with coworkers. This issue was highlighted in a study [7] published in Healthcare Informatics Research [8].

OCR also urged covered entities and business associates to regularly train staff members on cybersecurity issues. This can include but is not limited to employee training on phishing emails and when to report a cyber incident and to whom.

Employee security awareness was the greatest healthcare data security concern for 80 percent of surveyed health IT executives and professionals, a HIMSS Analytics survey found [8].
The September newsletter [8] also listed multi-factor authentication as a basic healthcare cybersecurity measure. This will usually include a password and additional security measures, such as a thumbprint or key card, OCR said. “A username and password may not be adequate to protect sensitive information, privileged accounts, or information accessed remotely,” OCR wrote. “As part of its risk analysis, an entity should determine what authentication schemes to use to protect its systems and sensitive information (e.g. e-PHI).”

Organizations should also ensure that they are regularly updating and patching their systems, the newsletter continued. These often fix critical security vulnerabilities and failing to keep tabs on the latest available updates could leave networks – and potentially ePHI – exposed to attackers.

Security patches are especially critical for ransomware prevention measures, a fact that has also been highlighted by the ECRI Institute [8].

2. FAIR4Health scenarios related to cybersecurity

Security-related factors that should be considered in the scenarios are:

❖ Type of data:
  ➢ Raw data, including personal identifiers.
  ➢ FAIR data, which could be one of the following:
    ■ Pseudonymised: personal identifiers are replaced with artificial identifiers, or pseudonyms.
    ■ De-identified: personal identifiers are prevented from being connected with information.

❖ Type of access:
  ➢ Through PPDDM (no human interaction with data)
  ➢ Through P2P
  ➢ Local access

❖ Users’ roles:
  ➢ Data managers (DM)
  ➢ Health researchers (HR)
  ➢ Data Scientists (DS)/Software Developers
  ➢ Healthcare professionals (HP)

The development of the scenarios is based on the FAIR4Health open community and its foreseen information communication exchanges.

2.1. Scenario 1: Local access to raw data

In this scenario, users (data managers, DM) access locally to raw data stored in their local facilities to process it and make it FAIR by using the FAIR4Health local agent. Users will perform processes of pseudonymisation/de-identification, curation, validation, mapping to
standard vocabularies, authoring and preservation of data in order to enable its sharing, reusability and actionability. In this scenario, only metadata is sent outside the users’ facilities to enable its discoverability.

![Diagram](image1)

**Fig. 1:** Scenario 1: Local access to raw data

### 2.2. Scenario 2: Local access to FAIR data

In this scenario, users (health researchers, HR) access locally to FAIR data stored in their local facilities to perform activities related with the research, such as updating their datasets or performing statistical analyses on them. In this scenario, only metadata is sent outside the users’ facilities to enable its discoverability.

![Diagram](image2)

**Fig. 2:** Scenario 2: Local access to FAIR data

### 2.3. Scenario 3: PPDDM access to FAIR data

In this scenario, users (data scientists, DS) trigger the execution of PPDDM methods to access locally to FAIR data stored in the data owner’s facilities to perform algorithmic operations on these datasets. In this scenario, the outcomes of these algorithmic operations will be sent to the FAIR4Health platform, which will be located outside the data owner's facilities.
2.4. Scenario 4: P2P access to FAIR data

In this scenario, once the users (health researchers, HR) have received the permission from the data owner, they will access to remote FAIR datasets through the FAIR4Health platform. In this case, FAIRified datasets (pseudonymised or de-identified) will be sent to the user, which will be located outside the data owner’s facilities.

2.5. Scenario 5: Access to eHealth services

In this scenario, users (healthcare professionals, HP) will access to eHealth services based on FAIR data through the FAIR4Health platform during their daily work. The following is an example of this scenario: During a routine consultation, a medical doctor wants to make...
use of the “Prediction of 30-days readmission risk” eHealth service. After checking that his patient fulfills the inclusion and exclusion criteria, he/she completes the online form provided by the FAIR4Health platform with the required clinical observations and launches the execution of the eHealth service, receiving after a few seconds the 30-days readmission risk prediction according to the entered variables.

![Fig. 5: Scenario 5: Access to eHealth services](image)

### 3. FAIR4Health Security requirements

This section provides an overview of relevant security requirements for the FAIR4Health system. The mindmap summarizes related topics on security requirements that need to be considered in the proposed approach. Based on these security requirements further technical recommendations are proposed, which are described in the next section of this document.

![Fig 6: Mindmap of general security requirements (blue) and specific security requirements (orange) for FAIR4Health](image)

The stated security requirements were derived according to the defined user scenarios and general security concepts established for shared, cloud platforms. We specified requirements that apply to all scenarios as well as specific ones for each scenario. Further
component-specific requirements are defined according to the components (FAIR4Health local agent, FAIR4Health Platform) from the use case scenarios.

Based on the use cases stated above and general cloud-related security requirements, FAIR4Health security requirements were derived, which can be grouped according to two classes:

1. **General Security Requirements**: Those are security requirements that are not specific to a certain use case nor FAIR4Health system component but concern the complete FAIR4Health system.

2. **Specific Security Requirements**: These relate to security requirements that matter especially for a specific use case or component.

### 3.1. General Security requirements

The following general security requirements are relevant to all identified FAIR4Health scenarios and the overall idea of the FAIR4Health approach:

**Identification & authentication**: Each actor and each component (data sources, PPDDM services, etc.) needs to be unambiguously identified within the FAIR4Health environment. The identification needs to be unambiguous, immutable and transparent.

**Logging**: is a fundamental feature for enabling confidentiality and security. For traceable actions, each triggered event within a software system should be logged with related information [9]. Considering the FAIR4Health approach, the following parts are considered crucial for logging:

- **Who triggered the event?** - Identification of the actor/service, who triggered the event. Requirements comprise a unique identity for each actor/service and related metadata.

- **When was the event triggered?** - The exact date and time when the event happened. This can be realized through an immutable timestamp, accessible from all actors.

- **What caused to trigger the event?** - The description of the event in the sense of unambiguously identifying what task/action has triggered the event. In this case it is necessary to determine a suitable way on how to denominate and identify events (e.g. severity levels similar to syslog, meaningful messages, for relevant cases even interdependency among events). It might be further useful to state any effect on the further process caused by the triggered event, e.g. deny the access for the use or depending on the severity send an email notification to administrators or other responsible actors.

- **Why was an event triggered?** - Specifies the context of the actual event and describes its intention. E.g. an agreement between data providing party and data retrieving party is necessary depending on the actual access on the data. A log event might be related to such an agreement for its completeness. Further the log might link interrelated events or create a history sequence of events based on the perspective (e.g. events triggered for the process of accessing FAIR data on the
platform for one user). Beside the proactive detection of security vulnerabilities, it might also be used for intrusion detection. It might be further used to verify the access.

Apart from the actual logging task, it is necessary to manage the gathered logs for secondary use (debugging) and ensure access authority as well as data integrity. Responsible roles or policies need to ensure confidentiality.

Data transmission and storage: Data transmission can occur over a point-to-point or multipoint channel. This raises the risk of data loss and manipulation during access on FAIR data. Data transmission over the internet needs state of the art protection for both, the endpoints as well as the transmission channel. The following requirements need to be met:
- Encryption of data needs to be ensured.
- Secure end-to-end communication enables that only the communicating participants can read the content of the messages.
- Data validation should be performed before transmission to avoid platform jeopardizing data inputs.
- A suitable backup plan needs to be part of the system including the data emerging throughout the system life cycle.

Endpoint-security: The underlying and dependent components of the platform need to be kept in a secure state as well. This comprises the security management of the different system levels like application layer (e.g. malware protection), operating system or virtualization (e.g. trust on cloud services/systems) and hardware protection (e.g. firewalls).

Authorisation is the required function of specifying access rights/privileges to data. It is necessary to secure 1) the intellectual property of the data set or the data mining algorithm and 2) to avoid data loss and privacy risks. In order to establish authorisation mechanism, two major concepts are well known: The first is called attribute-based access control and defines policies, which describe rules based on (user-)attributes for access control (Binary “If-Then”-Logic). The second concept is called role-based access control and defines user roles with specific access rules (e.g. user-role “healthcare professional” is allowed to access metadata of all data sets stored in the FAIR4Health platform). The management of authorisation and access control needs to be well defined and mandatory for all participating actors. The rules need to be elaborated in a joint-effort and transparently communicated according to the used concepts:
- It is required, to define/manage/enforce Service and User roles.
- It is required, to define/manage/enforce access rules and policies.
- Authorization needs to support different levels of granularity and security management.
  - General right management defines rules which are valid for the platform in general.
  - Feature specific right management defines rules for concrete specific elements of the FAIR data platform (e.g. special access rules for birth dates within a dataset).
Security Standards are highly relevant for the specific security requirements, yet general concepts contained need to be addressed. Well-known security standards comprise:

- ISO/IEC 27001
- ISO/IEC 27002
- ISO 27799

They not only describe security requirements limited to IT, but also deal with the integration of information security into the organizational structure and processes. The standard ISO/IEC 27001 also advocates the proven concept of the PDCA-cycle, comprising the steps Plan, Do, Check and Act to manage IT security for a continuous, frequent and quality-oriented improvement of information security.

3.2. Specific security requirements

3.2.1. Requirements related to Scenario 1: Local access to raw data

The Fair4Health Local Agent needs to ensure de-identification based on specified authorization policies. The policies can be general policies or policies that only apply to specific features. Variables affected by De-identification are called Quasi-identifiers, these can be for example dates like birthdates as well as gender and location. A common criterion to measure de-identification is k-anonymity, which means that each dataset is similar to at least k-1 other datasets on the identifying variables [10].

3.2.2. Requirements related to Scenario 2: Local access to FAIR data

In addition to the general logging, the Fair4Health Local Agent needs to ensure that the data is properly versioned in case of any modification. For data versioning a tradeoff between storage and recreation time has to be considered: The more storage used, the easier it is to recreate or access specific versions. In order to reduce storage, it is an established practice to save the differences of sequential depended data instead of full files (delta encoding). This approach opens up two contrary optimization problems, minimizing the storage size and minimizing the recreation time needed. Depending on the requirements the focus has to be set on optimizing for one problem over the other [11].

3.2.3. Requirements related to Scenario 3: PPDDM access to FAIR data

The trust in PPDDM or any other services that access FAIR data has to be ensured. This can be achieved by the following measures:

- The identification of a PPDDM service has to be unambiguous, immutable and transparent.
- Services must authenticate themselves via certificates signed by trusted CAs.
- A trusted party has to verify that the service meets the necessary requirements regarding data security and privacy/de-identification.
- The execution of PPDDM services has to be sandboxed in order to guarantee that the services can only access permitted resources.

3.2.4. Requirements related to Scenario 4: P2P access to FAIR data
Considering the general security requirements, no special security requirements could be identified for this scenario. However, endpoint-security as well as secure data transmission is crucial and need to be considered with high priority for this use case.

3.2.5. Requirements related to Scenario 5: Access to eHealth services

For this use case, security has to be adapted to existing health IT technologies, e.g., through IHE. As stated in the general requirements, it is relevant to ensure authorisation of access and manage potential data access agreements among providing party and the accessing healthcare organisation or healthcare professional.

3.2.6. FAIR4Health Local Agent Security Requirements

The FAIR4Health Local Agent acts as the direct interface for local raw data and the shared FAIR4Health platform. Therefore, the following security requirements need to be met:

❖ Trust has to be established between the local agent and the FAIR4Health platform (remote access) and between the local agent and local clients (local access)
❖ Fair4Health Local Agent needs to ensure secure and reliable data source connection.
❖ Integrity of the Fair4Health local agent needs to be ensured
  ➢ Containerization can be used to ensure that the local agent has a consistent environment that can be securely ported to other systems without manipulation.
  ➢ Certification can further confirm confidence about the Agent’s purpose and origin.
❖ The local agent needs to monitor and log data access which might be used by security systems to detect malicious activities by deviant behaviours.

3.2.7. FAIR4Health Platform Security Requirements

The FAIR4Health Platform is responsible for the complete management and realization of the FAIR4Health data sharing. Apart from the general security requirements, the following aspects need to be considered:

❖ The platform needs to detect malicious access and react accordingly: The behaviour of participating parties needs to be monitored, so that suspicious behaviour can be detected and blocked.
❖ FAIR4Health Platform needs to ensure the secure connection to the related actors and to the local agents.

4. Technical recommendations related to scenarios identified

4.1. General Concerns

This report recommends utilizing/implementing open source protocols and standards while designing the cybersecurity related components of FAIR4Health. Security related leaks
should be monitored continuously, this can be achieved through automated security tests in production environment. A dedicated cybersecurity interface can be implemented for the end-users so that they can always see a summary of the ongoing security related activities and the implemented precautions. This interface can also be designed as a security wizard so that the respective administrative users can manage the security related activities by turning modules/features on and off based on specific scenarios and requirements.

4.2. Identity Management Module

This module should be responsible for the unique identification of all kind of clients within the FAIR4Health system. Clients can be the users (Data Manager, Health Researcher, Data Scientist, Healthcare Professional) of the several components of FAIR4Health while the components themselves can be the clients interacting within the FAIR4Health system.

❖ All clients should be managed (CRUD operations).
❖ Non-managed clients cannot make any transactions within FAIR4Health. Authentication & Authorization Module should ensure this by interacting with this Identity Management Module.

Object Identifiers (OIDs) [12] are an identifier mechanism standardized by the International Telecommunications Union (ITU) and ISO/IEC for naming any object, concept, or "thing" with a globally unambiguous persistent name. OIDs can be used as a basis for the identification of the clients and components within this module.

Patient identities which exist in the raw data should be handled within a different mechanism than the client identification. If our scenarios require patient identification across different data sources, then we need a patient identification module. The following standards can be used to design the patient identification module:

❖ IHE PIX: The Patient Identifier Cross Referencing (PIX) Integration Profile [13] supports the cross-referencing of patient identifiers from multiple Patient Identifier Domains. This profile enables a secure and trusted identification of patients although they have different identifiers within different data sources.
❖ HL7 FHIR MPI: A Master Patient Index (MPI) [14] is a service used to manage patient identification in a context where multiple patient databases exist. Healthcare applications and middleware use the MPI to match patients between the databases, and to store new patient details as they are encountered. MPIs are highly specialized applications, often tailored extensively to the institution's particular mix of patients. MPIs can also be run on a regional and national basis.

4.3. Authentication & Authorization Module (Auth Module)

This module should provide standards-based authentication and authorization mechanism for all clients of the FAIR4Health system. The following standards and profiles can be used during the design of this module:
Improving Health Research in EU through FAIR Data

- Authentication and authorization should follow well-established international standards. OpenID Connect [15] is a simple identity layer on top of the OAuth 2.0 [16] protocol and seems to be a good candidate to be integrated in this module. It allows Clients to verify the identity of the End-User based on the authentication performed by an Authorization Server, as well as to obtain basic profile information about the End-User in an interoperable and REST-like manner. This functionality should interact with the Identity Management Module for the user/client profiles.

- SMART Health IT [17] is an open, standards-based technology platform that enables innovators to create apps that seamlessly and securely run across the healthcare system. Using an electronic health record (EHR) system or data warehouse that supports the SMART standard, patients, doctors, and healthcare practitioners can draw on this library of apps to improve clinical care, research, and public health.

- SMART on FHIR [18] is a set of open specifications to integrate apps with Electronic Health Records, portals, Health Information Exchanges, and other Health IT systems. This set of specifications exposes an authorization mechanism named as SMART App Authorization Guide to authorize the applications within the ecosystem. This specification is lately published by HL7 itself with the name of SMART App Launch Framework [19].

4.4. Audit Module

Logging requirement can be considered in two different levels. One for the software logging that needs to be done in different levels within the source code to track the activities with info messages, inform the associated parties with error messages and debug the system when necessary. On the other hand, on the application integration layer, each transaction (i.e. data access, query request) is very important and needs to be logged explicitly for further user inspection. The following standards and profiles can be used while designing this module:

- IHE’s The Audit Trail and Node Authentication (ATNA) Integration Profile [20] establishes security measures which, together with the Security Policy and Procedures, provide patient information confidentiality, data integrity and user accountability. ATNA contributes to access control by limiting network access between nodes and limiting access to each node to authorized users. This profile not only serves to the logging mechanism, it also provides facilities for user and connection authentication. The logging facility is provided by the Audit Trail. The Audit Trail needs to allow a security officer in an institution to audit activities, to assess compliance with a secure domain’s policies, to detect instances of non-compliant behavior, and to facilitate detection of improper creation, access, modification and deletion of Protected Health Information (PHI).

- HL7 FHIR (Audit Event) [21] provides a means to maintain security logs. FHIR repository to be used as the audit trail repository, the audit messages are designed to be FHIR resources with this standard. A record of an event made for purposes of maintaining a security log. Typical uses include detection of intrusion attempts and monitoring for inappropriate usage.
4.5. PPDDM Security Module

Existing PPDDM approaches can be categorized under two concepts when the security is in question:

1. De-identified data leaves the data source so that the PPDDM engine on the platform can process whole data for predictive modelling. Only software components access data. Once the processing completes, data is deleted permanently. Communication is performed through end-to-end encryption.

2. Encrypted data leaves the data source. PPDDM engine processes the encrypted data and tries to build the models by using the encrypted data. No client can decrypt the data except the data owner. This is also called homomorphic encryption in literature. This approach is rather new in theory and no practical implementation exists today for real world use.

Blockchain concept can be analyzed as a candidate to be utilized during the design and implementation of the PPDDM security module. Smart contracts can reflect and link the agreements made among the peers to the transactions occurring through PPDDM.

4.6. End-to-end Security Module

Asymmetric encryption algorithms should be used to maximize the level of encryption of the transmissions among components. On the other hand, FAIR4Health is about sharing data between computer systems. Hence, respective developments in the secure data sharing arena should be explored. The following standards and profiles can be used while designing this module:

- Public-key cryptography or asymmetric cryptography, is a cryptographic system that uses pairs of keys: public keys which may be disseminated widely, and private keys which are known only to the owner. In such a system, any client can encrypt a message using the receiver’s public key, but that encrypted message can only be decrypted with the receiver’s private key. Established Public-key concepts are used for a variety of internet standards, including Secure Socket Layer (SSL), Transport Layer Security (TLS), Pretty Good Privacy (PGP) or GNU Privacy Guard (GPG).

- Dat [22] is a secure peer-to-peer protocol for sharing data between computers. Dat can be considered as a candidate to be used as the base protocol while sharing data between remote FAIR4Health components.

4.7. Utilization of the Cybersecurity modules within the scenarios

This section provides the placement of the cybersecurity modules in each identified scenario. The modules should provide the identified security functionalities:

- Identity Management Module
  ➢ Identities of all clients are managed.
  ➢ Trust is ensured.

- Auth Module
  ➢ Clients are authenticated (in line with the identity management)
  ➢ Authorization is handled through roles, scopes and authorization rules
Improving Health Research in EU through FAIR Data

- Audit Module
  - All transactions are audited into an audit repository.
- End-to-end Security Module
  - All transactions are encrypted using asymmetric encryption algorithms.
- PPDDM Security Module
  - Based on the selected data mining method, this module should ensure that only allowed components process the FAIR data.

4.7.1. Scenario 1: Local access to raw data

4.7.2. Scenario 2: Local access to FAIR data
4.7.3. Scenario 3: PPDDM access to FAIR data

Fig. 9: Security Modules attached to Scenario 3

4.7.4. Scenario 4: P2P access to FAIR data

Fig. 10: Security Modules attached to Scenario 4
4.7.5. Scenario 5: Access to eHealth services

![Diagram showing Audit Module, FAIR4Health Platform, End-to-end Security Module, Auth Module, Identity Management Module.]

**Fig. 11**: Security Modules attached to Scenario 5

5. Conclusions

In this section, the findings are translated into both technical and functional recommendations related to cybersecurity issues to serve as inputs for FAIR4Health project deliverables D2.1 “Technical recommendations for the FAIRification tool implementation” and D2.2 “Functional design of the FAIR4Health platform and agents - FAIRification workflow”, respectively.

5.1. Technical and functional recommendations

**Security-Rec #1:**
- Functional recommendation: **All users, data subjects and software components need to be managed by an Identity Management Module.**
- Technical recommendations:
  - OIDs can be used as a basis for the identification of users.
  - Data subjects (patients) could be managed by either IHE PIX profile or HL7 FHIR MPI if needed

**Security-Rec #2:**
- Functional recommendation: **All users need to authenticate and receive authorization** for the different functionalities available in the FAIR4Health platform.
- Technical recommendations: Several standards could fit this purpose:
  - OpenID Connect on top the OAuth 2.0 protocol is proposed for this task.
SMART Health IT
SMART on FHIR/SMART App Launch Framework

Security-Rec #3:
❖ Functional recommendation: **All operations and transactions performed by both software and users over the datasets must be logged and audited.**
❖ Technical recommendations: Several standards could fit this purpose:
   ➢ IHE Audit Trail and Node Authentication (ATNA) profile
   ➢ HL7 FHIR (Audit Event)

Security-Rec #4:
❖ Functional recommendation: **When software components access de-identified data the communication must be encrypted**
❖ Technical recommendation:
   ➢ Blockchain concept is a potential candidate to implement this communication flow.
   ➢ Public-key cryptography systems (e.g. GPG, PGP, TLS, SSL)

Security-Rec #5:
❖ Functional recommendation: **Asymmetric encryption algorithms should be used to maximize the level of encryption of the transmission among components.**
❖ Technical recommendations: Several standards could fit this purpose:
   ➢ Public-key cryptography systems (e.g. GPG, PGP, TLS, SSL)
   ➢ Dat standard for secure P2P data sharing

References

Improving Health Research in EU through FAIR Data

[16] OAuth 2.0, https://oauth.net/2/
[18] SMART on FHIR http://docs.smarthealthit.org/