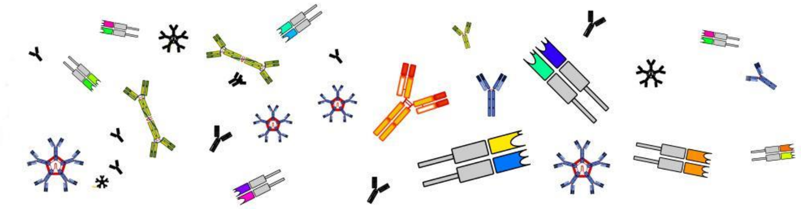


Deliverable 3.1

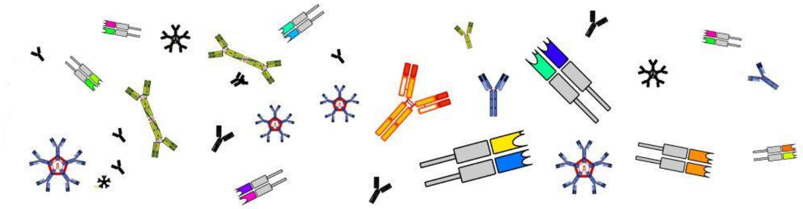
Holistic Security and Privacy Concept

Work Package Number: WP3

Work Package Title: Layered Data Security

Type: Report

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)



Document Information

iReceptor Plus Project Information	
Project full title	Architecture and Tools for the Query of Antibody and T-cell Receptor Sequencing Data Repositories for Enabling Improved Personalized Medicine and Immunotherapy
Project acronym	iReceptor Plus
Grant agreement number	825821
Project coordinator	Prof. Gur Yaari
Project start date and duration	1st January, 2019, 48 months
Project website	http://www.ireceptor-plus.com

Deliverable Information	
Work package number	WP3
Work package title	Layered Data Security
Deliverable number	D3.1
Deliverable title	Holistic Security and Privacy Concept
Description	Establish a holistic approach to address the security and privacy aspects in iReceptor Plus, resulting from the Ethical, Legal and Social Implications (ELSI) that need to be considered generally in Genetic and Genomic Research and in particular about AIRR-seq data.
Lead beneficiary	INESC TEC
Lead Author(s)	INESC TEC
Contributor(s)	Ascora, SFU, BIU,Sorbonne, UTSW, Haifa, APHP, Mitmynid
Revision number	

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Revision Date	
Status (Final (F), Draft (D), Revised Draft (RV))	D
Dissemination level (Public (PU), Restricted to other program participants (PP), Restricted to a group specified by the consortium (RE), Confidential for consortium members only (CO))	PU

Document History

Revision	**Date**	**Modification**	**Author**
	15-04-2019	Consolidated version circulated internally	Artur Rocha, Ademar Aguiar, José Alexandre Teixeira
	17-04-2019 to 23-04-2019	Contributions on data types, providers and consumers in iReceptor Plus	Brian Corrie
	24-04-2019, 25-04-2019	Document revision	Artur Rocha, Brian, Corrie, Tobias Hinz
	28-04-2019	Document revision	Christian Busse
	29-04-2019	Document revision	Felix Breden, José Alexandre Teixeira
	30-04-2019	Document revision	Artur Rocha, Gur Yaari

Approvals

	Name	**Organisation**	**Date**	**Signature (initials)**
Coordinator	Prof. Gur Yaari	Bar Ilan University	30 Apr, 2019	GY
WP Leaders	Artur Rocha	INESC TEC	30 Apr, 2019	AR

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

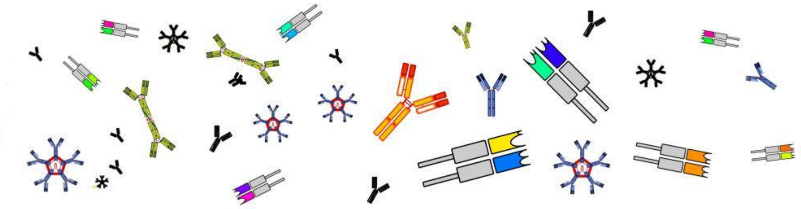


Table of Contents

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Executive Summary

The purpose of this deliverable is to establish a holistic approach to address the security and privacy aspects in iReceptor Plus, resulting from the Ethical, Legal and Social Implications (ELSI) that need to be considered generally in Genetic and Genomic Research and in particular about AIRR-seq data.

Many AIRR-seq data sets are from human subjects, and as such are subject to strict confidentiality and security constraints. In many cases, biopharmaceutical companies are producing some of the most significant AIRR-seq data. It is challenging for these companies to integrate their data into any implementation of an AIRR-seq Data Commons unless they can protect their data in terms of security and licensing.

This deliverable addresses the concepts, approaches and possible solutions for the data security layer to be implemented in the iReceptor Plus platform to fully enable the secure sharing of data.

The deliverable describes a set of guidelines with standards and approaches to implement layered security between the components of the platform, providing multiple levels of authentication, authorization and auditing.

These security layers will provide mechanisms for data stewards to implement control at different levels of granularity. Access may be restricted at multiple levels (e.g. through authentication) to specific types of data (e.g. through role-based authorization) and according to its privacy level. Furthermore, adequate monitoring and auditing mechanisms shall be provided for the previously identified access layers (e.g. through logging and audit controls, including the use of blockchain).

Deliverable description

Deliverable D3.1 is the main result of the work done in WP3 Task 3.1.

The goal of this task is to create a holistic concept about addressing security and privacy aspects of the project, ensuring that the concept is applicable to and followed by all iReceptor Plus components, considering existing approaches and both the general requirements of the health domain and the specific requirements around AIRR-seq data.

Secure data sharing will be enabled by implementing layered security between the individual components across the iReceptor Platform, enabling data stewards to control access to their confidential data while at the same time using the iReceptor Platform to enable the secure sharing of those data where permissible through formal agreements such as Ethics Board approvals and Material Transfer Agreements.

A layered security model will provide multiple levels of authentication and authorization. These security layers will provide mechanisms for data stewards to implement levels of control at different levels of granularity, restricting access at multiple levels (through authentication) as well as restricting access to specific types of data (through role-based authorization) and tracking data access and use (e.g. using logging or auditing technologies via blockchain technology).

In addition to each layer providing control over access to data at different levels of granularity, a layered security model also helps protect data, requiring an attacker to compromise each layer of the security platform in order to access data. Security layers for a data repository include control over access at the client level (the tool or system trying to access data), the user level (the user for which the client is requesting access), and the data level (control over, and tracking of which data a user has access to within the repository). This task will follow the implicated requirements formulated in Work Package (WP) 11.

The results of this task are two-fold: firstly, a set of guidelines is defined which will define the standards and approaches to be used within the project in order to use a high security and privacy approach.

Secondly, the task will provide concrete requirements to WP1, Task 3.2 in WP3 and WP2 to WP7 on what those work packages need to follow in order to stay compliant to the concept elaborated by this task.

1. Introduction

The National Institute of Standards and Technology (NIST) defines security as:

"The protection afforded to an automated information system in order to attain the applicable objectives of preserving the Integrity, Availability, and Confidentiality of information system resources [1]."

In general, security is achieved through specific mechanisms, such as encryption, signature, authentication, access control; and security architecture to ensure confidentiality, integrity, and availability.

Layered security describes the practice of combining multiple mitigating security controls, such as authentication steps to protect resources and data. The main goal of a layered approach to security is to make sure that a breach or failure of one layer does not compromise the entire system.

Holistic Security

Holistic security is an approach with the goal to integrate all the system's components of an organization and safeguard them. Thus, the purpose of holistic security is to provide continuous protection across several levels.

This approach comes with the principle of creating several layers of security measures in order to reduce the risk that a threat becomes a hazard, by action of the mechanisms implemented in each successive layer. These defenses are usually of different nature so that any weaknesses that one line of defense may have does not easily allow a risk to materialize since other defenses also exist thus preventing a single point of weakness. For every category of threat, there should be an effective control deployed to mitigate the threat.

Multi-layered Security

There are multiple layers of security considered important when designing and implementing solid and effective security mechanisms to protect an information system. Even though there is not a one-size-fits-all solution to handle cyber security, the following layers are commonly considered.

- **Human Layer.** The first, and hardest, layer of security to control is the human layer, since people do not always behave as security experts would. Effective security programs combine policies, norms for acceptable behavior, education, and controls to mandate specific behaviors.
- **Physical Layer.** The second layer aims to protect the physical premises where the data and system is housed. The goal is to prevent unauthorized people from entering and physically accessing the hardware storing your data.

- **Endpoints Layer.** The third layer concerns running unauthorized programs that could compromise data assets, such as malicious software logging keystrokes, stealing data, causing a machine to crash, or encrypting data to make it unusable.
- **Network Layer.** The fourth layer controls who has digital access to the data and from where, using firewalls, virtual private networks, intrusion detection and protection systems to ensure that unauthorized individuals cannot connect to the data.
- **Application Layer.** The fifth layer ensures that each application exposing data is properly managing access to the data and cannot be compromised. To limit the potential for vulnerabilities, we can use techniques such as: active penetration testing, vulnerability scanning, and source code analysis.
- **Data Layer.** The last layer is devoted to add protections around the data itself, including identity and access management to control who accesses the data and what they can access.

Since the first five layers are usually deal with at an organizational level by the service provider, this deliverable focuses mainly on the mechanisms to be implemented in the scope of the Data Layer by the iReceptor Plus software platform.

Data Security Layer

To implement a data security layer, we must first define the data and its components that need a higher level of protection. User-permission management can be used to control who can view, update, create, and delete data of certain levels of security and sensitivity. Authorization can be used to verify who a person is to provide them the appropriate level of access rights. These fundamental mechanisms of identity and access management become essential in a data protection layer.

In addition, the data within the system can be secured at multiple layers. For example, the data can be stored encrypted or the data itself can have an additional layer of encryption to ensure that even when accessed directly, people cannot do anything with the data without the right keys.

Outside the strict perspective of a security mechanism to prevent actual threats to the system it can also be used to restrict the authorization given to a user to a certain extent. This means that the operations that imply the access to more restrict data will in turn require more authentication mechanisms. Thus, each authentication mechanism can be abstracted to a single layer.

Materializing this concept to an information system, we could grant any user with the basic access to browse data that has no confidentiality constraint. However, the user may want to retrieve from the system data that is more sensitive thus requiring them to authenticate to prove to the system they have privileges to access it. This can be applied to a *n-tier* layered system with each level requiring a higher form of authentication, for example a Multifactor authentication.

2. Regulations and Standards

Information security is an important topic in today's Internet-based systems. In terms of social requirements, usually depicted in the legal specifications, security and privacy of citizens' medical records are sensitive issues.

Most of the legislation entitles the citizen as the owner of their own medical data and grants them the highest right on their own medical records. This means that it is mandatory to get a consent from the person (data subject) when their medical records are accessed for whatever purpose. Hence the reason it is one of the frequently debated issues in all sectors of activity: political, medical, legislative and industrial.

Other than the primary use of medical data for the treatment of patients, the medical data can also be used for secondary purposes. These secondary purposes can be involved in medical research, survey or academics. This is important for knowledge contribution for research on the medical data to know more about diseases, medicines and sociological factors. This results in improvements of medicines, medical practices and technologies.

In the context of iReceptor Plus, we highlight a set of regulations, standards and policies that should be considered in terms of conformity when designing and implementing iReceptor Plus components and systems, such as: General Data Protection Regulation[1] (GDPR), Ethical, Legal and Social Implications[2] (ELSI), and European Open Science Cloud[3] (EOSC).

- **GDPR**. The European General Data Protection Regulation is a regulation in EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA).
- **ELSI**. The US National Institutes of Health (NIH) and the Department of Energy (DOE) have recognized the need to prepare for the social impacts of the Human Genome Project, and they have created a program for studying its ethical, legal, and social implications.
- **EOSC**. The European Open Science Cloud is a European Commission project to provide a public data repository conforming to open science values. The EOSC-hub has been streamlining security policies to be shared across infrastructures, including a harmonised Acceptable Use Policy (AUP) and a GDPR policy framework to help tackle Data Privacy issues. The EOSC-hub provides a set of services for research support, including services for sensitive data[4].

[1] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[2] https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-92-2620-11

[3] https://eosc-portal.eu

[4] https://www.eosc-hub.eu/services

3. General Data Protection Regulation

The European General Data Protection Regulation[5] (GDPR) is a regulation in EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA). It also addresses the export of personal data outside the EU and EEA areas.

The GDPR aims primarily to give control to individuals over their personal data and to simplify the regulatory environment for international businesses by unifying the regulation within the EU.

Definitions

Understanding the dispositions of the GDPR requires clear definitions about topics it regulates. The following list is a subset of the definitions included in Article 4[6] of (EU) 2016/679.

- **Personal data** is any information relating to an identified or identifiable natural person (data subject). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
- **Processing** is any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
- **Restriction of processing** consists on marking stored personal data with the aim of limiting their processing in the future.
- **Profiling** is any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.
- **Pseudonymisation** consists of personal data processing in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

[5] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[6] https://gdpr-info.eu/art-4-gdpr/

- **Filing system** is any structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis.
- **Controller** is an entity that determines the purposes and means of the processing of personal data.
- **Processor** is an entity that processes personal data on behalf of the controller.
- **Recipient** is an entity to which the personal data are disclosed, whether a third party or not. Public authorities which may receive personal data in the framework of a particular inquiry in accordance with Union or Member State law shall not be regarded as recipients; the processing of those data by those public authorities shall be in compliance with the applicable data protection rules according to the purposes of the processing.
- **Third party** is an entity other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data.
- **Consent** of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.
- **Personal data breach** is breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.
- **Genetic data** is personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question.
- **Biometric data** is personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data.
- **Data concerning health** is personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.

Data Subject Rights

The following list summarizes the rights of the data subject[7], which are listed in articles 12 to 23 of the GDPR.

Right of access. The right of access encompases 1) the right to know whether data concerning the data subject are being processed and 2) if so, access it (GDPR Article 15).

[7] https://gdpr-info.eu/chapter-3/

Right to rectification. When personal data is inaccurate, controllers need to correct them (GDPR Article 16).

Right of erasure or right to be forgotten. Upon request of the data subject, providing that there are no legal grounds to keep it. It includes with additional stipulations, among others if personal data has been made public (GDPR Article 17).

Right to restriction of processing. The right of the data subject to limit the processing of their personal data along with several rules and exceptions (GDPR Article 18).

Right to be informed. The right to be informed states that for personal data that have undergone an action as a consequence the above mentioned data subject rights (GDPR Articles 16, 17 and 18), the controller must inform recipients who got these data, where feasible. Additionally the data subject also has a right to know who are the recipients who got their data (GDPR Article 19).

Right to data portability. Data subjects have the right to receive their personal data, which they have provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided (GDPR Article 20).

Right to object. Data subjects can state they do not want the personal data processing to be done or going on. Data subjects can, within specific conditions, exercise the right to object and the right to be forgotten (GDPR Article 21).

Right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her (GDPR Article 22).

GDPR Articles 13 and 14 establish the **information to be provided by the controller to the data subject**, respectively in the cases where personal and non-personal data have been collected.

Furthermore, GDPR Article 23 establishes restrictions to the rights and obligations of the data controller or processor due to EU or member state laws.

Security requirements for GDPR compliance

In summary, complying to the GDPR and in particular to Article 25, the so-called secured by design, requires abiding to the following set of principles:

- **Fairness**. Process personal data lawfully and transparently by informing users where and when data processing is taking place, and the description matches the process.
- **Have a reason**. GDPR states that personal data should be obtained only for specified, explicit and legitimate purposes.

- **Minimize data**. Personal data should only be kept if it is relevant and strictly limited to what is necessary in relation to processing purposes.
- **Stay updated**. Personal data must be up to date and accurate.
- **Keep only as long as necessary**. Data should be removed when it is no longer required, there should be given the consent to store and process data, and ensure data is portable.
- **Process appropriately**. Personal data should be processed in an appropriate manner to prevent loss, damage or destruction.

4. Security Requirements and Mechanisms

When designing and implementing software systems we should consider a broad range of aspects. One of these aspects is *security* and it is one of the most important aspects of a system, particularly when it is composed of smaller, loosely coupled subsystems, as established by good practices in software design.

In a service-oriented architecture environment, like the iReceptor Plus Platform, the platform is composed of services from different sources, such as iReceptor and VDJServer. The application executing at the different organizations are integrated but not tightly-coupled. Also, the composition is technology-agnostic, as the applications are not based on similar technologies, languages or platforms.

This distributed aspect of iReceptor Plus and the technology-heterogeneity of the integrated applications raise challenges for security. The inter-organizational workflow among the different parts execute in a decentralized manner. These pipelines process sensitive information, which need to be secure in the local data stores of the organizations and also while they communicate with each other. Moreover, these organizations would also like to make sure that only authorized users with specified and previously agreed permissions should get access. This is the organizational aspect of security.

In a simplified way, there are two main types of security requirements to consider: the security of the service itself, composed by parts; and the security of the contents, both as sources and results (see Figure 1). These requirements and related security mechanisms are briefly overviewed in the following sections.

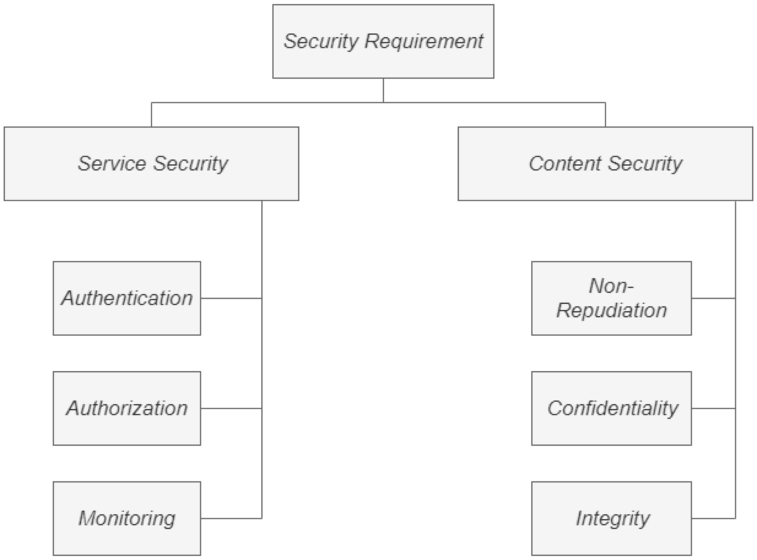


Figure 1. Service and content security requirements.

This project is funded by the European Union's H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

Identification, Authentication and Authorization

Identification is the ability to identify uniquely a user of a system or an application that is running in the system. *Authentication* is the ability to prove that a user or application is genuinely who that person or what that application claims to be. *Authorization* protects critical resources in a system by limiting access only to authorized users and their applications. It prevents the unauthorized use of a resource or the use of a resource in an unauthorized manner.

An authorized user may belong to one of the organizations and has their own unique identity and their own *role* in the context of the application. For this matter the *Identification* of a user is one of the basic security requirements. This requires that all the organizations involved should have defined policies for the identification of users, i.e. for their *authentication*. These policies specify the type of *credentials* accepted by the organization, the security mechanism that is being applied and the algorithms and information about who will perform the validation of the credentials provided. Additionally, the organizations should have *authorization* policies defining the permissions that each role has. These permissions specify what each role is allowed to access once the user is authenticated.

The core security mechanisms can be seen as platform-independent. For instance, *authentication* can be refined to *direct authentication*, *brokered authentication*, *single sign-on* (SSO) or *identity federation* (see Figure 2). Likewise, *authorization* can be *refined to role-based access control* (RBAC), *context-based access control* (CBAC), and *attribute-based access control* (ABAC).

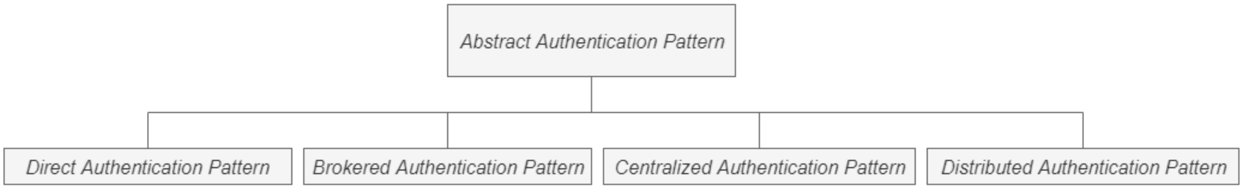

Abstract Authentication Pattern
Direct Authentication Pattern
Brokered Authentication Pattern
Centralized Authentication Pattern
Distributed Authentication Pattern

Figure 2. Possible refinements for Authentication: Direct, Brokered, Centralized or Distributed.

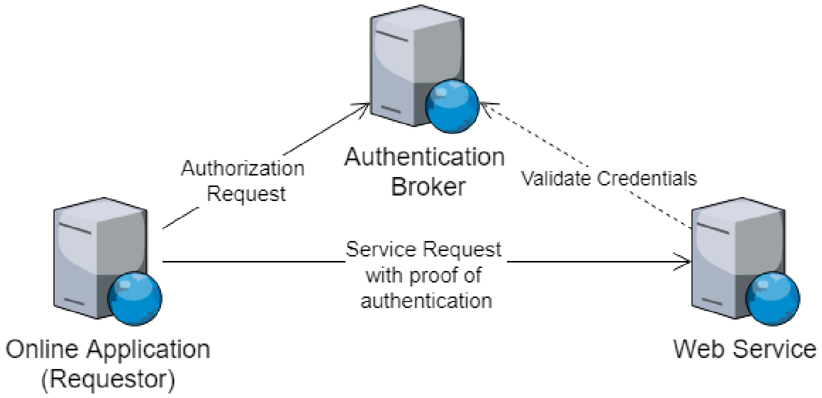


Figure 3. Example of *brokered authentication*.

Figure 3 represents a high-level view of a *brokered* authentication. For instance, while the security token is always verified, the service typically does not need to interact with the broker to perform the verification. The reason for this is that the token itself can contain proof of a relationship with the broker, which can be used by the service to verify the token.

Identity Validation

The identity validation is performed through *authentication*, whereas the role/permission verification is performed through *authorization*. The security mechanism at the partner's organization validates the user's credentials representing his identity for granting them access to the resources. In this case, the requested organization has a security policy for authentication of an internal user. The policy defines the accepted credentials, that can be either a digital certificate or a username-password pair.

Two-Factor Authentication

Two-factor authentication (2FA) is one of the most effective ways to protect personal data and private information shared online and uphold the principle of *integrity and confidentiality* enshrined in the new GDPR regulation, which requires personal data to be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.[8]"

This is advisable whenever there is the need to send, receive or access sensitive data such as health records. This sensitive data is often shared across devices through platforms like Google and other companies that have integrated 2FA functionality into their login processes.

[8] https://gdpr.algolia.com/gdpr-article-5

2FA is a system that consists in using two different, but related methods to identify a person. It is a more secure method of identification than a reusable password. By combining the random number of a token with a private PIN, the resulting passcode provides more trustworthy user authentication.

These two methods fit in one of the three categories:

- Something that only the user knows, such as a password
- Something that is in the user's possession, such as a mobile phone
- Something that is the user's immutable property, such as biometric features like a fingerprint, or retina scan.

Using a password and a PIN would not satisfy the 2FA since it uses two methods in the same category. However, in the event of authenticating in a web application the user can supply the password and a code generated by a physical authenticator, or having the code be sent to his mobile phone.

For example, when a person withdraws money from an ATM they need to know the PIN and have a physical card to complete the transaction.

Since 2FA is able to recognize devices, so using the same phone or computer, a physical resource, will often provide on itself the second factor of authentication without adding an extra interaction for the user. This intends to make the authentication process more difficult for third parties that may have malicious or unauthorized access to the system.

There is a common use of One-Time passwords (OTPs) and Public Key Infrastructure (PKI) to secure applications and data.

One-Time Passwords (OTP) and 2FA

One-Time passwords are a single use password that is generated specifically for one login session on a computer. For an user to authenticate in the system, this password has to be delivered to an external source such as a text message to the user's mobile phone. Despite the fact that OTP can solely be used for authentication, it is an improvement to a 2FA setup where the user can use his ID, password and the OTP to access the system.

Direct and Cross-domain authentication

Direct authentication happens when the authentication mechanism resides in the same application domain. Whenever it should be necessary to communicate with an external source, such as the communication between VDJServer and iReceptor data repositories, this requires a Cross-domain Authentication in which a user has to get authenticated by an application of the external domain. To access the services from an external domain some more additional credentials have to be supplied such as a *Security Token*. The cross-domain authentication problems can be solved with *Brokered Authentication* and *Identity Federation*. Ultimately to apply the same security requirements between platforms/organizations/repositories different security patterns are required in these different scenarios.

The underlying requirements for the iReceptor Plus will inevitably lead to interaction between more than one partner organization. Users from these organizations may access resources from within the organization and from other organizations. In order to access these resources the user has to prove his identity, which will enable him to get an appropriate role in the system to grant them the required permissions that a specific role has access to.

Role-based Access Control (RBAC)

Access management is concerned with managing who has access to specific things or processes in a systems environment.

The system development cannot be confined to select a few users who can address all the functional and non-functional aspects of systems. The tasks are more commonly distributed among roles according to their competence. Thus, different roles communicate with each other using system models and domain specific languages.

The Role-based Access Control model is one of the most common approaches. Its inherent logic is in the form of if-clause statements and generally a few generic roles can cover most of the use cases. For example a user of type *researcher* can do operations on a particular set of data whereas a user of type *clinician* has only permissions to explore data.

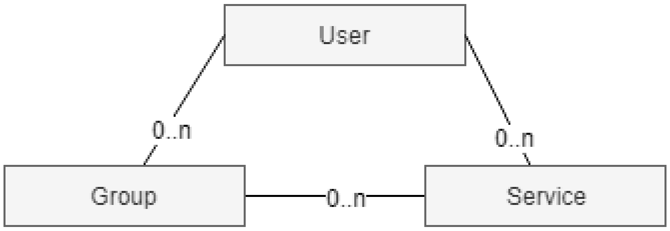


Figure 4. Example of Role-based Access Control.

However, RBAC tends to be functionally insensitive. Roles are strict, meaning that a user gets access to a resource regardless of what they are doing and why. When different types of data exist, with some of that data being *sensitive data,* a more nuanced and contextually sensitive control system can help ensure that everyone has access to all of the data they need, but *only* the data they need.

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

To some degree, this can be mitigated by defining increasingly more granular roles and *ad hoc* roles with limited memberships to fit emergent needs. However, this solution quickly grows overwhelming due to constant maintenance of these rigid roles.

Attribute-based Access Control (ABAC)

An alternative to the RBAC model is the attribute-based access control (ABAC) model. Attributes are effectively anything that can define a user, their environment, or operational conditions. As the name implies, ABAC assigns permissions based on attribute criterias.

ABAC systems control access with rules that define how access logic assess and responds to selected attributes. Comparing to roles, that are collections of permissions to resources, rules define conditions for permission assignments. Therefore, rule writers can pick which attributes and conditions are important in a given process. Moreover, rules can change depending on how the system responds. This gives ABAC a lot of flexibility that can be crucial in dynamic business and development environments.

Non-repudiation

To hold a user accountable for his actions it is also necessary to ensure accountability in the systems regarding access to the data or services. This is achieved through non-repudiation.

This security requirement asserts that a user should not be able to deny having participated in a particular interaction. Meaning that when data is requested from one of the data repositories the user should not be able to deny having to perform the query on that data. Therefore, this security requirement is necessary for accountability and auditing, becoming an essential requirement used for verification of access and usage of system resources.

Non-repudiation can also be performed using different patterns depending on the scenario. There can be voluntary non-repudiation that is most applicable if the agent/user belongs to the system/platform. Considering the iReceptor Plus project requirements it is likely that intercommunicating but independent subsystems will exist, therefore a Trusted Third Party will be involved to ensure non-repudiation across the system constituent parts.

Confidentiality

The *confidentiality* service protects sensitive information from unauthorized disclosure.

When sensitive data is stored locally, access control mechanisms might be sufficient to protect it on the assumption that the data cannot be read if it cannot be accessed. Encryption of data is preferable if a higher level of security is required.

However, sensitive data that is transmitted over a network, especially over the Internet as it is an insecure network, access control mechanisms are not effective against attempts to intercept data such as wiretapping.

Data Integrity

The *data integrity* is a concept that detects whether there has been unauthorized modification of data.

Data can be tampered in two ways: accidentally through hardware and transmission errors or a deliberate action in the form of an attack. Regardless, many hardware and transmission protocols already have mechanisms that detect and correct these transmission errors. Thus, the purpose of data integrity is to detect deliberate actions.

Additionally, data integrity is intended to detect whether data has been tampered, it does not provide a mechanism to restore data to its original state if it has been modified.

Access control mechanisms can contribute to data integrity as data cannot be modified if access is denied. But, these access control mechanisms are not effective in not secure networking environments, such as the Internet.

Security Monitoring

Security monitoring[9], sometimes referred to as *security information monitoring* (SIM) or *security event monitoring* (SEM), involves collecting and analysing information to detect suspicious behavior or unauthorised system changes on your network, defining which types of behavior should trigger alerts, and taking action on alerts as needed.

Personal Data in Logs

Under GDPR, access logs, error logs and security audit logs will now be considered to hold personal information. Thus, it is necessary to protect the IP address and cookie data as they would be personal identifiers. Also, personal data cannot be collected or stored without documentation of individual consent. Nonetheless personal data **can** be collected and stored with web server logs to help detect and prevent fraud or unauthorized system access.

[9] https://www.hpe.com/emea_europe/en/what-is/security-monitoring.html

Blockchain and GDPR-compliance

Blockchain technology has already gained significant interest in several industries including healthcare, supply chain management and the financial industry. While the exact use cases vary from industry to industry, it is a common statement that blockchain offers a new, secure way of storing and processing large volumes of data.

However, the immutable nature of the blockchain technology is a wall to GDPR regulations. Decentralized systems are still a gray area in terms of legislature. Current provisions demand that users are allowed to remove or correct their personal data.

A decentralized ecosystem assumes to a single controller that could do so. No individual can delete data recorded on a public blockchain, thus compromising the “right to be forgotten” rule. Actually, no authority can effectively control a public blockchain. The core concept of this technology is to eliminate single-source ownership and return the data rights back to its users. Removing data from a private blockchain is technically possible, though challenging.

To address this issue, there are already solutions that take on a different approach. Instead of trying to erase data from the blockchain, the relevant decryption keys necessary to decode certain entries would be deleted, thus rendering the data unobtainable since it is no longer possible to be decrypted.

The aforementioned solution leads to a scenario of usage in which instead of explicitly sharing their data with third parties, users would only give their permission to access the said information from the blockchain. Basically, only providing them the key to their data, and this key would be volatile enough so the users have the power to revoke the access at any time. Thus, the original data can not be changed or misinterpreted by the receiving party.

Factually, blockchain-based identity management assumes “privacy by design”, which is the biggest demand from the regulators.

Blockchain privacy

Taking the example of the traditional banking model, the banking models achieves a level of privacy by limiting access to information to the parties involved and the trusted third party. The need to announce all transactions publicly (inherent to public blockchains) precludes this method, but privacy can still be maintained by breaking the flow of information in another place, by keeping public keys anonymous. In the network, the public can see that someone is sending an amount to someone else, but without information linking the transaction to anyone. This is similar to the information released by stock exchanges, where the time and size of individual trades is made public, but the parties involved are not revealed.

5. Data, Data Producers, and Data Consumers in iReceptor Plus

The type of security requirements for a specific type of data is driven largely by the type of data, who produced the data (and its associated data privacy constraints), and who consumes the data (and their associate roles and access rights). The challenge for implementing a security policy and picking a specific security approach for iReceptor Plus is mapping these three dimensions to a technical solution. In order to simplify this mapping, we enumerate the types of data, data producers, and data consumers in the iReceptor Plus community and ecosystem. It should be pointed out that many of the definitions and classifications that are provided here are driven by the early work that is currently underway in the WP 1 Use Case development task (Task 1.2, Deliverable 1.2)

Types of Data

There are four types of fundamental, high-level data that the iReceptor Plus ecosystem needs to consider:

- **Study Metadata**: This is the metadata about a given research study. It typically consists of data about the study, the subjects in the study, and the biological samples (and how the samples were obtained and processed) from those subjects. In the context of iReceptor Plus, the structure and form of this data is well understood and is guided by the AIRR Community's recommendations for the Minimal Standard (MiAIRR[10]) for storing and sharing AIRR-seq data. *This is a fundamental data component of the iReceptor Plus Platform and this data is stored in the repositories in the AIRR Data Commons and must be protected appropriately*.
- **AIRR-seq Data**: Each biological sample above is processed and eventually sequenced to produce a set of sequence data. This sequence data is a statistical sampling of the actual Adaptive Immune Response Repertoire of that individual at a specific time point. This data is then annotated with a set of features that are critical to understanding the Adaptive Immune Response. Combined (the sequences and their annotations), this data is known as AIRR-seq data. *This is a fundamental data component of the iReceptor Plus Platform and this data is stored in the AIRR Data Commons and must be protected appropriately*.
- **External Metadata**: This is metadata from outside of the iReceptor Plus Platform that we want to link to the core data components above. This might be data that is linked to Study Metadata (other immune relevant data such as microbiome, clinical, cytokines and chemokines data - WP6) or AIRR-seq data (such as a sequence's Epitope as stored in an external repository such as IEDB (iedb.org)). Implementing the data security and access restrictions for these external data sources is outside the scope of iReceptor Plus but it is necessary to manage access (through authentication and authorization if required) to these external data sources.

[10] https://www.nature.com/articles/ni.3873

- **Analysis Metadata**: This is data that is generated from the application of an Analysis Pipeline to the combination of a set of Study Metadata, AIRR-seq Data, and External Metadata. Because the output of an analysis pipeline produces data about other data, we term this Analysis Metadata. Analysis tools can produce metadata that is associated with either Study Metadata or AIRR-seq data. One of the key development efforts in the iReceptor Plus project is to expand the analysis capabilities of the iReceptor Plus platform. At this point in the project, it is unclear as to whether the Analysis Metadata will be stored and managed by the platform of if the analysis results will be provided to (and be the responsibility of) the user. *If the Analysis Metadata is stored and managed by the iReceptor Plus platform, then this data will be treated as a fundamental data component of the platform and will be protected.*

Data Access Levels

Each of the data elements discussed above can be classified to have a specific data access level, based on how the data was produced, who the data steward is, and the ethics and privacy constraints under which the data was acquired. In the remainder of this document we use the following data access classifications:

- Public - no restrictions to what can be done with the data. Data can leave the platform
- Open - no restrictions to what can be done with the data, access to data needs to have an audit trail to determine who had access. Data can leave the platform and once it does traceability is lost. Possibly used on data that was already public before included in the platform. Access to data summaries and metadata could also be monitored for profiling reasons.
- Protected Open - access only to those that have authorization, once granted access data can leave the platform. The user is responsible for adhering to data privacy protocols once the data has left the platform (can download, individual is responsible).
- Protected Platform - access only to those that have authorization, data does not leave the platform but individuals can view some or all individual data elements (no external download, but data is moved between protected systems for processing and analysis).
- Private Platform - access only to those that have authorization, data does not leave the data providers system, but individuals can view some or all individual data elements (no external download, data remains on and is processed by the host system).
- Private - access to only those that have authorization, no actual data is available to users, only summary statistics and analyses are available.

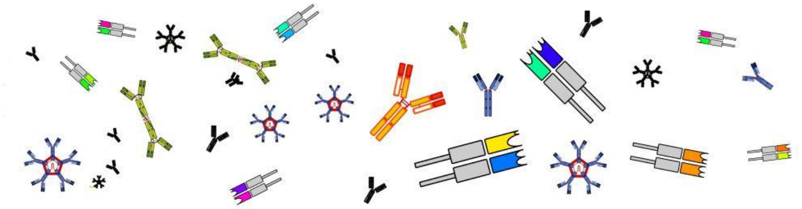
Data Providers

The type of individual or organization that produces data can also be classified. Although some classes of organizations would typically have a more restricted access policy (e.g clinics, industry), it is possible (and common) to have data of all access levels to be produced by each class of organizations. Some indicative

This project is funded by the European Union’s H2020 Research and Innovation Programme under Grant Agreement No. 825821 and Canadian Institutes of Health Research (CIHR)

suggestions of the access restrictions that might be common for a certain type of organization are given below, but this list should not be considered complete. Classes of data providers are:

- Individual researchers: An individual researcher (and their students) might perform one (or more) study(s) involving Study Metadata and AIRR-seq data with the goal of publishing the research results. Given that most research publications either require or encourage the publication of appropriately pseudo-anonymized research data, such data would often be *Public* or *Open*. In some instances, the research data produced may fall under a more restrictive ethics or data privacy agreement, meaning data would be *Protected* or *Private*.
- Research groups: A group of researchers (and their students) within an institution might perform a larger number of studies involving Study Metadata and AIRR-seq data with the goal of publishing the research results. Given that most research publications either require or encourage the publication of appropriately pseudo-anonymized research data, such data would often be *Public* or *Open*. In some instances, the research data produced may fall under a more restrictive ethics or data privacy agreement, meaning data would be *Protected* or *Private*.
- Research consortia: A group of researchers (and their students) across several institutions might perform a larger number of studies involving Study Metadata and AIRR-seq data with the goal of publishing the research results. Given that most research publications either require or encourage the publication of appropriately pseudo-anonymized research data, such data would often be *Public* or *Open*. In some instances, the research data produced may fall under a more restrictive ethics or data privacy agreement, meaning data would be *Protected* or *Private*.
- Industry (service providers) : A service provider that performs sequencing and/or sequence analysis as a service on behalf of a client (either industry or academic) might produce Study Metadata and AIRR-seq data. In this instance, the service provider's privacy and access restrictions to the data would be mandated by that of the client. As such, a specific data might take on any of the access levels given above.
- Industry (bio-pharma): A bio-pharma company might be performing research into the immune response to develop improved medical treatments and new drugs. Such research, and in particular turning that research into a viable product, is time consuming, costly, and of high commercial value.
  - These companies may produce large AIRR-seq and related data sets that are of high value to the company and therefore need to be either *Protected* or *Private*.
  - At the same time, research that leads to drug design can benefit from comparing *Protected* or *Private* data to *Open* or *Public* data.
  - In addition, some companies value Open Data and are willing to publish some of their data for the use of the general research community as *Open* or *Public* data.
- Hospitals/Clinicians:
  - A hospital or clinic may collect AIRR-seq data to prescribe treatment or monitor response to treatment. This data would be clinical data and would therefore be *Private*.



  - Comparing clinical AIRR-seq data to other AIRR-seq research data may help to inform patient treatment. As such, a clinician may want to compare their clinical *Private* data to *Open* or *Public* data.
  - In addition, in the case of a research hospital, where allowed by ethics and privacy, some data produced by clinicians in a hospital may allow for the *Open* or *Public* sharing of that data.

Data Consumers

There are a number of types of potential users for data within the iReceptor Plus Platform (data consumers). Below is a high-level list and some possible classes of data they might access. Given that all data consumers have access to *Open* and *Public* data, these classes apply to all data consumers and therefore are not discriminated below:

1. Individuals who have provided data to a research study: Access to *Private* data collected from that individual.
2. Patient in a clinical setting: Access to *Private* data collected from that patient.
3. Individuals who are interested in AIRR-seq data (citizen science): *Open* and *Public* data only.
4. Individual researchers: *Protected* or *Private* data that the individual is authorized to access through their roles in the research project based on the project's data sharing agreement.
5. Research groups: *Protected* or *Private* data that individuals or groups are authorized to access through their roles in the research project based on the project's data sharing agreement.
6. Other research platforms: *Protected* and *Private* data that either other platforms (software tools or applications) or users from other platforms are authorized to access through the platform's or the individual's role in the research project based on the project's data sharing agreement
7. Industry (service providers): *Protected* or *Private* data (both data internal to the company, data from the service provider's clients, and external research data from partners) that individuals or groups are authorized to access through their roles in the company or research project based on the company's and/or client's and/or project's data sharing agreement.
8. Industry (bio-pharma): *Protected* or *Private* data (both internal to the company and external research data from partners) that individuals or groups are authorized to access through their roles in the company or research project based on the company's and/or project's data sharing agreement.
9. Clinicians: *Protected* or *Private* data (both internal to the hospital and external research data from partners) that individuals or groups are authorized to access through their roles in the hospital or research project based on the hospital's and/or project's data sharing agreement.

6. Security in iReceptor Plus

System Architecture

Tracking data as it moves across a network is rarely a simple task. It is likely that an audit of many networks would reveal sensitive personal data tucked away in places that no one would ever expect to find it, stored unprotected in applications and databases across the network.

A possible first approach is to analyse the system from a technical perspective and identify all the points and places where sensitive data is processed, transmitted and stored, including the data flows into and out of numerous applications and systems. It is precisely this flow that needs to be the focus of a holistic approach. Not only the platforms where data is being transmitted to are important also the environment support where they are being transmitted is also critical.

For iReceptor Plus, figure 5 depicts the current understanding of the information flows occurring across the several components of the system. Based on the identified information flows, the next section includes a proposal of the security layers and respective supporting mechanisms that are candidates for use in the implementation of those security layers.

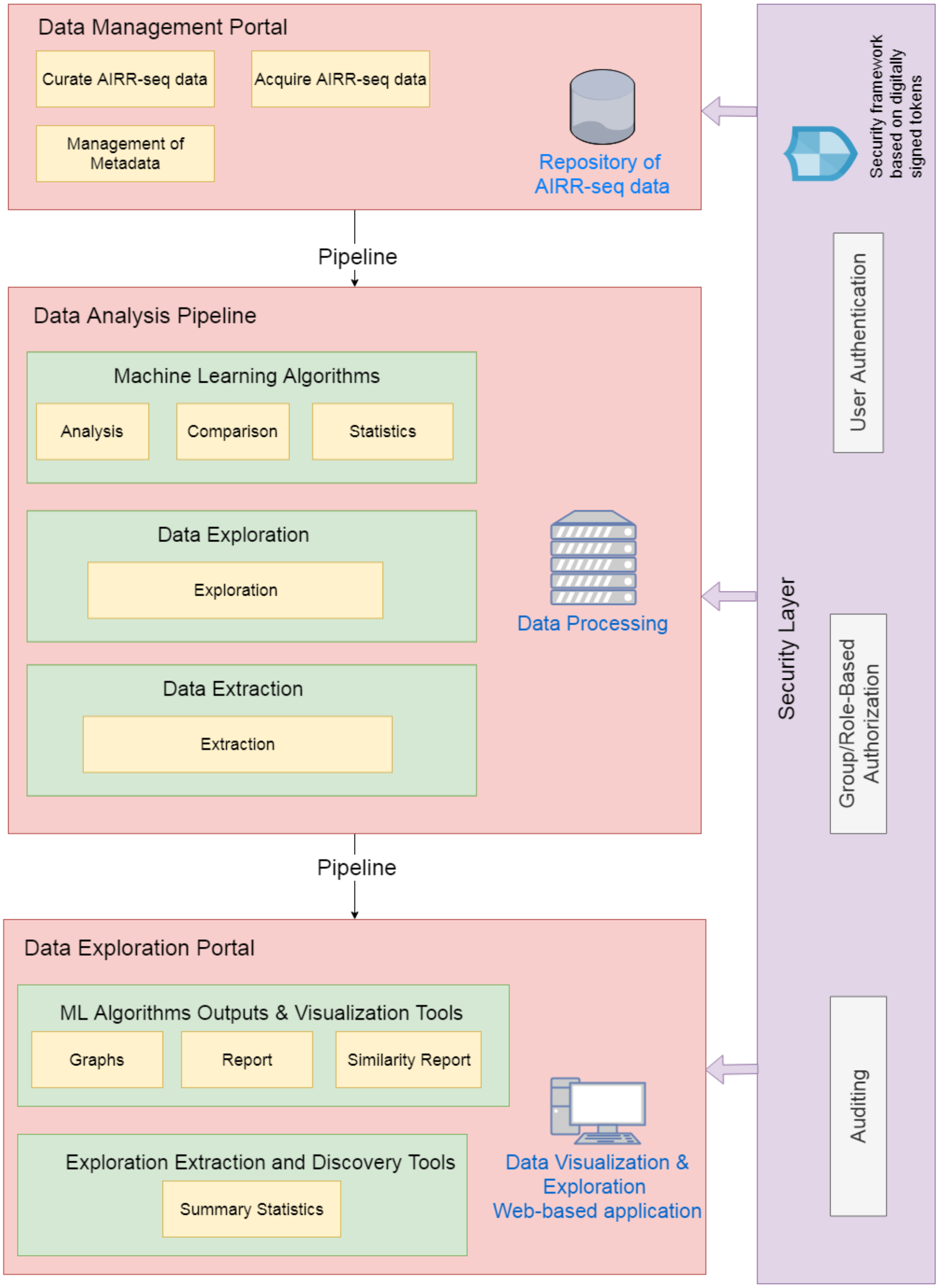


Figure 5. Information Flows in iReceptor Plus.

iReceptor Plus Data Security Layers

Bearing in mind the previously considered information (data flows, data types, data producers, and data consumers) in the context of iReceptor Plus, several layers of data security have been identified, and for each layer, a set of possible implementations of different security mechanisms for authentication, authorization, auditing and monitoring. The rationale behind the differentiation of the proposed layers resides mainly in the granularity, sensitivity, and type of use of the data being browsed and retrieved from the repositories. The main outcomes of the remaining implementation tasks in Work Package 3 will revolve around refining, designing, and implementing these security layers.

Data Type/Layer	Purpose	Security Mechanisms
Public	**Any data consumer** interested in browsing *Public* Study Metadata, AIRR-seq Data, External Metadata, or Analysis Metadata. **Download allowed**.	Basic: - No authentication and authorization. - Anonymous logging (e.g. using browser fingerprint)
Open	**Registered data consumers** interested in performing searches and browsing over *Open* Study Metadata, AIRR-seq Data, External Metadata, or Analysis Metadata. **Download allowed.**	Basic: - Role Based Access Control (RBAC). - Logged access (e.g. using user logs)
Protected Platform	**Registered data consumers** interested in performing searches and browsing over *Protected* Study Metadata, AIRR-seq Data, External Metadata, or Analysis Metadata. **No download, data visibility, data can be moved between protected systems within the platform for processing and analysis.**	Basic: - RBAC - Logged access (e.g. using user logs) - Logged data movement (e.g. using data transfer logs) Advanced: - Two Factor Authentication (2FA) - RBAC - Robust logged access (e.g. using blockchain) - Digital Rights Management (e.g. encryption)
Private Platform	**Registered data consumers** interested in performing searches and browsing over *Private* Study Metadata, AIRR-seq Data, External Metadata, or Analysis Metadata. **No download, data visibility, data remains on and is processed by the data provider's system.**	Basic: - RBAC - Logged access (e.g. using user logs) Advanced: - Two Factor Authentication (2FA) - RBAC - Robust logged access (e.g. using blockchain)

Private	**Registered data consumers** interested in performing searches and browsing over *Private* Study Metadata, AIRR-seq Data, External Metadata, or Analysis Metadata. **No download, no data visibility (summary statistics only), data remains on and is processed by the data provider's system.**	Basic: - RBAC - Logged access (e.g. using user logs) Advanced: - Two Factor Authentication (2FA) - RBAC - Robust logged access (e.g. using blockchain)

Given the above, it is important to point out that:

- Any Data Producer (e.g. Researcher, Clinician, Industry) might produce any Data Type (*Public, Open, Protected, Private*).
- It is possible for any Data Consumer, through following appropriate processes and policies, to be assigned a role that would authorize them to access *Private* data. That is, in the context of the iReceptor Plus project, almost all users need roles and all Data Types, with the exception of *Public* data, require role based access control.
- The level of access that a Data Consumer has for a specific piece of data (or data set) is based on a combination of the Data Type (*Public, Open, Protected, Private*), the Data Providers who produced that data, the roles that the Data Consumer has approved for access to the data set, and the Data Consumer's assigned roles.
- To secure the systems and the data, the authentication mechanism used may range from no-authentication to two-factor authentication, according to content sensitivity.
- Authorization will be based on Role-based Access Control mechanisms. Sensitive data may require additional mechanisms such as the ones used for digital rights management (DRM) or encryption.
- The auditing and monitoring of the access will range from normal logging of access to robust logging and auditability, possibly using blockchain.

The above data security layers and the example security mechanisms that are listed provide Work Package 3, and indeed all of the Work Packages within the iReceptor Plus Project, with the necessary foundation to begin the design and implementation of these security layers. These security layers will need to be implemented across many of the components within the iReceptor Plus Ecosystem, including at the user interface of the "Data Exploration Portals" (WP1, Task 1.3, 1.4, 1.5), at the Data Repository level (WP3, Task 3.2.1), at the Web API level (WP3, Task 3.2.2), at the Analysis Tool level (WP4, Task 4.4).

7. Conclusions

This document provides a holistic view of the security and privacy concepts that underlie the iReceptor Plus project and platform. It introduces the concept of layered security and discusses relevant regulations and standards (including the GDPR). It then discusses data security requirements and their mechanisms for implementation. Lastly, it describes the data requirements for the iReceptor Plus Platform (Data Types, Data Producers, and Data Consumers) followed by a description of the architectural components of the iReceptor Plus Platform on which data will reside. As such, this document provides the data security foundations that will allow the iReceptor Plus team, through the other tasks in Work Package 3, to design and develop a layered security framework (Task 3.2) at the core of the iReceptor Plus Platform as well as providing a foundation for auditing and the prevention of data manipulation (Task 3.3). In addition, this document, combined with the deliverables from Work Package 3 on monitoring the security aspects of the project in the context of the GDPR (Task 3.4), the Ethics Work Package (WP11) and its deliverables, and the Data Management Plan for the project (WP10, Deliverable 10.2), will provide a baseline for data security for all Work Packages within the project. This holistic view will ensure that the design, implementation, and monitoring of data security will be performed as an intricate and fundamental part of the project.