Masaryk University
Faculty of Informatics

Optimisation of user digital identity gathering process

Master's Thesis

Bc. Dominik František Bučík

Brno, Spring 2021

Masaryk University
Faculty of Informatics

Optimisation of user digital identity gathering process

Master's Thesis

Bc. Dominik František Bučík

Brno, Spring 2021

This is where a copy of the official signed thesis assignment and a copy of the Statement of an Author is located in the printed version of the document.

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Bc. Dominik František Bučík

Advisor: RNDr. Michal Procházka, Ph.D.

Acknowledgements

I want to thank RNDr. Michal Procházka, Ph.D., for his supervising and valuable advice. I also want to thank my colleagues from CESNET and ICS MU, who were very willing to help and their encouragement when writing my thesis.

Abstract

Digital identity is an important concept used in a digital world. It is a base for enabling processes of identification, authentication, and authorization. An essential aspect of the digital identity is the level of trust it expresses. In an internationally provided research infrastructures, the identity has to be strongly verified. However, instead of creating a new identity with performing verification tasks, the infrastructure can rely on already existing user identities.
In this thesis, we analyze the digital identity and concept of its trust level. We focus on its usage in the environment of international life science infrastructures. We discuss different types of digital identities, their strong and weak points, and their applicability in the chosen domain. We also analyze the process of associating the external identity with user representation in the target environment, both from its owner and consumer perspective. We inspect an already existing infrastructure and evaluate the used approach. The primary result of this work is a set of guidelines that will help the reader to select the correct type of external digital identities the environment can rely on, what communication protocols to use, how to build the environment, how to design the association process, and what documentation to provide for users and relying services.

Keywords

digital identity, level of trust, attributes, access management, external identity, eID, digital certificate, academic identity federation

Contents

Introduction

We can observe digitalization having a huge impact on our lives. Many tasks can now be accomplished using online tools. A several of them need the context of identity of the person who is performing them. For such purposes, the digital identity concept has been developed and is now heavily used.
We can observe this trend in the research area as well. Groups of people performing research have joined their efforts together and established communities around their field of interest. One such is the life science research community. We can observe many projects being established, funded, and supported by governments. To cooperate or perform their regular tasks, many researchers need to use specialized tools or perform highly resource-consuming tasks, like high-performance computing, storing an enormous amount of data, or just access to a specific data set. Instead of getting such resources locally, the communities often establish common infrastructures where the participants can share and provide such things. Such a situation can happen even on an international scale.
To be able to control access to such resources, particular persons have to be identified in the digital environment. This allows the communities not only to control the access but also make people responsible for their actions, apply billing on consumed resources, or create new cooperations with less effort. Also, it creates a great place to look for what tools and resources the researchers need.
Particularly in the life science field, the researchers might need access to sensitive data, especially genetic material, disease analysis, or clinical records. For such purposes, the identity in the digital world needs to be strongly verified. However, due to the international scale of such infrastructures, this task can become challenging to solve. Such infrastructures often rely on already existing digital identities instead of creating new ones.
In our thesis, we look at the concept of digital identity, and primarily on the representation of identity verification. We also discuss the level of trust that can be put into the identity, its representation, and the specific requirements on digital identity in the context of international infrastructure operated for purposes of life science researchers. We

analyze different types of digital identities the users might already have, their attributes, and their applicability in the mentioned context. After that, we take a closer look at the process of integrating the best possible options, both from the identity owner and its consumer perspective. We discuss the usable protocols for communication while exchanging information about the person's identity. We also analyze a sample infrastructure and look at particular aspects connected to the digital identity. At the end of our thesis, we focus on providing guidelines for designing and implementing an environment satisfying the requirements we have set for such a case.

1 Digital identity

In recent years the world has been moving towards an online way of doing things. This trend has been initiated mainly by the evolution of modern technologies and becoming more available and popular. A great example is the business sector. Nowadays, almost every subject presents its services or products (at least a subset of them) in the digital world. It serves both as a showcase area and a channel for selling the products to the consumers. The business subject usually wants to know who is the potential customer. In other way, they need to identify the client. We illustrate such situation on an online store example. Its owner has a clear intention - to sell the offered items to the customers. However, each client has their own needs or preferences, and the seller needs to identify them. Such an environment creates an explicit requirement to have a customer representation. It can be just a database entry storing the person's name and billing address. To prevent conflicting representations, an unique identifier might be needed, e.g., an artificial number given to each customer. Clearly, such an entry creates a mapping of the real-world person to the particular context in the digital environment.
Another field that enforced the development of representing a person in the computer world has been the military. From the start of using digital technologies, there has been a need to perform access management tasks As the digital technologies became more and more involved in this field, it has also moved towards the digital way of expressing identity. The particular subjects started identifying themselves using digital technologies. As a result, an automated system could then perform decisions based on the subject's identity stored digitally. For example, imagine the famous scenes from the movies where a special agent unlocks a secret chamber by scanning the fingerprint or iris. Only if the agent were allowed to enter it, the door would open. Such methods did not remain just science-fiction scenes and became a reality. Not only that, they grew into everyday tasks in ordinary people's lives (e.g., scanning fingerprint on the smartphone to unlock it).
These trends of identifying the involved subjects using digital technologies can be observed in many more places. In each area, the need

for the data stored about the person can differ. For this reason, having just a single representation of a person's identity is not enough. Each environment has its unique set of requirements on the semantics, syntax and what data needs to be available. In some of them, capturing the relationships among different subjects might be beneficial as well. In conclusion, we can see the people nowadays need to be identified in the digital world. The context where the identification is needed influences what data needs to be stored, if we want to know relationships between different subjects, and also how precise the information needs to be. Such data sets have evolved into a concept of a digital identity.

1.1 Definition of the digital identity

Digital identity can be defined in many ways. We can describe it is as a set of data about the real-world subject. This information aims to identify and distinguish a person or a thing in a virtual environment. For this thesis's scope, we will consider only human being's identities and omit the latter. The data needs to describe the person uniquely to make the identification possible. This data set is often referred to as identity attributes. The amount of the information stored can differ depending on the environment in which the identity resides. Data representation can vary as well, and even the semantics of the attributes can change. So far, no global solution that would serve all environments exists. An attempt to solve this problem has been the design and implementation of Public Key Infrastructures (PKIs). However, this attempt has partially failed due to technical, economic, and social reasons. [1]
To better illustrate why the number of data changes for each context, let us imagine a system where users communicate with each other. In such a case, the identifying data might be just a nickname. Users that want to communicate could exchange these identifiers and contact each other. On the other hand, in a university information system, the identity can consist of identification number, name, university email, and birth date. In the state public registry maintained by the government, the digital identity can contain a unique person identifier (such as a birth number), the person's full name, the person's

picture, address of residence, passport number, and much more. As we can already see from the examples given, the digital identity tends to contain more precise and larger information set as the environment becomes more formal and the need for the exact person identification increases.
We have already mentioned in section 1 that identification data is not the only part of identity. As Philip J. Windley writes in his book Digital Identity [2], it also contains information about the subject's relationship to other entities. For example, in the mentioned university environment, we could state that the person is a student or employee. This little piece of information can provide us with a much bigger picture. Based on the subject being an employee, we can assume that they have got an office, use different parts of the information system, or access the parking garage.

1.1.1 Security and issues

Having an identity in the digital world provides many benefits. Unfortunately, it comes also with several difficulties that need to be solved. One of such is the protection of the identity itself. The system consuming the identity makes decisions based on the identity. If a person not represented by the identity gets access to it, the association becomes incorrect. As a result, the wrongly identified person might get access to resources unreachable for them or potentially reserved only for the identity owner. In a digital world, the attacker can make many harmful transactions in little time, and the damage can be potentially massive. Thus, securing the identity and access to it in the best possible manner is a crucial thing.
A typical attack on identity ownership in the digital world is impersonification, which we have briefly described in the previous paragraph. It is described as a situation when an attacker gets successfully identified as another entity. In other words, we can say that it gets access to (or uses) a different subject's identity. Protection against this kind of attack is a challenging thing. Solutions are often based on improving the authentication procedure, which we further described in section 1.2. For example, one solution is increasing trust in the identity claim demonstration or using more than one proof. However, the whole

process needs to consider users' points of view and not make using the system a problematic experience.
Another issue is the up-to-date correctness of identity attributes and propagation of value modifications. Some attributes can become outdated but still be presented as valid in places where they have not been modified yet. To make things even worse, we can find ourselves in a situation with no data validity procedure available at a particular moment for some of the attributes. An example could be the user's affiliation with an organization. Let us suppose that a person has registered into some system as an employee of an organization. Based on this affiliation, the user was granted access to specific parts of the system. After some time, the mentioned person has left the job. Without propagating the change to the end-service, it would incorrectly evaluate the person's rights, resulting in unauthorized access to the system's protected parts.

1.2 Identification, authentication, authorization and digital identity related processes

As already mentioned in section 1, the main reason for developing a digital identity concept was the need to identify a particular subject and perform access management tasks. These processes consist mainly of three parts: identifying the particular entity, verifying identity claim validity, and evaluating the entity's rights. In other words, the digital identity is the main driver for the ability to perform identification, authentication, and authorization.
First, the subject needs to present who or what it is. In the process, the subject claims that a particular digital identity represents it in the given context. An essential attribute in the process is an identifier of the subject's representation. It is usually a unique thing that distinguishes the identities in the whole environment. The uniqueness, however, might not extend across the borders of a single environment. The process of identification can, for example, consist of providing the identifier of an identity. It is important to note that the process of identification does not extend beyond this claim and does not involve any sort of verification or validation of the identity that we claim. [3]
Authentication is, in an information security sense, the set of methods

we use to establish a claim of identity as being true. [3] Usually, these tools include a process during which the subject persuades us about the authenticity of the identity statement. It can be carried out in several ways, often referred to as factors. A factor can be something the subject knows (secret password), something it is (biometric information), something it has got access to (device), something it does, or where it is. [3] In some scenarios, presenting more than just one factor might be needed. By using multiple factors, the claim about identity validity can become more robust.

Unfortunately, the identification and authentication do not provide any information about the subject's rights. As it has identified itself and proved the claim's correctness, the environment can extract additional data to decide the set of allowed actions. Resolving this set and verifying that a particular action is in the allowed actions set is called authorization. The process can be described on a simple web application, in which the users are categorized by granting a specific role - visitor and administrator. For instance, a person with the administrator's role might see agendas hidden to the webpage's regular visitor. Based on the user's role, after performing the authorization, the system can decide what data it needs to fetch and display appropriate application sections.

So far, we have considered digital identity to identify the person for authentication and authorization purposes. However, incorporating the representation of a person in the virtual environment enables providing additional security features. Firstly, we want to mention accountability. The NIST[1] defines this term as the principle that an individual is entrusted with safeguarding and controlling equipment, keying material, and information and is answerable to proper authority for the loss or misuse of that equipment or data. [4] Imagine a system that keeps an audit of actions performed by each user. As a result of having a way to identify the actor, we can map actions taken to a real-world person. Thus, the subject can be easily made accountable for the performed actions.

Another vital thing is non-repudiation. Jason Andress describes it as a way of preventing the sender from denying that they sent the message. [3] This concept can also expand to the activities performed

1. National Institute of Standards and Technology

in the digital environment. An application could keep a history of the performed actions and the information on who has been the actor. If a subject later states that they have not performed any actions, it can be easily verified and associated with the real actor.

1.3 Concept of an identity trust

An important attribute associated with identity is its trustfulness. This property expresses how reliable the identification is. Every identity-based decision made is dependent on this attribute. Decision making person or system must believe that the identification is valid and the identity attributes are correct. The other point of view is the trust of the identified person in the digital identity. The user must believe that the system has performed correct identification, which happens by authentication. During this process, the user usually has to present some confidential information that proves identity ownership. Due to this need, a belief in the system protecting confidential information and using it correctly is required.
When establishing a trusted digital identity, the person's real-world identity and attributes have to be verified. This requirement implies the need for a means that will support the confirmation process. For example, the person could have to come to a specific office of a company operating the system that needs to know its identity. An employee could ask for an identification card, verify and fill in the person's details into the system. With less strict verification requirements, another system could ask users to fill in the information without performing additional validity checks.
As evident from the previous paragraph, different contexts require different levels of identity trust. Currently, there is no unique way of classification. Each environment might set up its evaluation criteria and thus qualify the trust differently. An example can be the classification of digital signatures. By the eIDAS regulation, [5], there are three digital signature levels. The first and the weakest one in the trust hierarchy is called Simple digital signature. By definition, it can be anything in a digital form that the user uses for signing (e.g., a picture of a handwritten signature attached to an email). The next level is the Advanced digital signature. In this case, the signature has to be

created using a digital certificate. There are no further requirements for this level of trust. The top in the hierarchy is the Qualified digital signature. It requires a qualified digital certificate to be used and created using capable means for creating the signature. The qualified digital certificate is issued after going through a predefined process with strict identity verification.
Another example could be classifying the identity using REFEDS Assurance Framework. [6] In this case, the identity contains an attribute with one of the predefined values that express the trust level. The last example we want to mention is the definition and assurance levels established in the ISO/IEC 29115:2013 [7].

1.4 Digital identity in life science infrastructures

In section 1 we have declared that identity requirements and representation change depending on the context in which it is used. The information amount, syntax, and semantics of the attributes differ according to the particular environment's needs. Also, the minimal level of trust required for the correctness of the association represented by the identity changes as well.
Further in our thesis, we focus on the environment of Life Science (LS) communities. In several cases, such groups of people are established at the international level and consist of a potentially large number of involved organizations. Many of such communities are operating common infrastructures that support their needs in the computation field. Providing the researchers with tools, computational means, and sharing resources is the main reason for establishing such specialized information systems. They are built to provide and manage access to data as well as enable its further processing via the provided tools.
Such an environment establishes a vital requirement for identifying the participating subjects. As sometimes the data may contain sensitive information, the identification and access management processes are critical tasks. The infrastructure has to ensure that the identification is correct and the user's identity is verified. Therefore, the identification of the users has to achieve a certain level of trust. As the whole environment might be pretty dynamic in its matter of members, another critical thing is keeping the identity attributes up to date. As a

result, the platform must contain the user's identity and provide tools for authentication, authorization, and user management. A common solution to such a situation is establishing an AAI[2].
The AAI environment consists of two main participating subjects - the users and the (relying) services provided to the users. The goal of the users is to use the functionality offered by the relying services. Such a service might be a platform for executing data analysis tasks. However, the services may restrict provided resources based on the identity of the user. Therefore, they have to recognize the user to be able to apply the restriction rules correctly. In most cases, users already have an already existing digital identity. The AAIs often take advantage of this fact and rely on these existing user representations rather than creating new ones.
The infrastructure acts as a joint point for the users and services. From the user's point of view, it provides authentication tools and serves as a unified point for accessing the services. From a relying service perspective, the AAI holds the identification data, performs authentication, and provides authorization data. For instance, the users could log in to the relying service with an existing digital identity via the AAI. After successful authentication, the AAI performs authorization tasks. If all rules are met, it forwards the service an identifier of the user's identity and additional attributes that the service can further process.
AAI might often be a transparent component of the whole environment, both from the user or relying service perspective. Nonetheless, it plays an essential role in the ecosystem. Apart from the tools and functionality mentioned so far, it might even serve as a solution for ethical or legal aspects. The AAI can enforce an agreement with legal documents like terms of service. For example, it can show the user a webpage where the users have to mark that they have read the document and confirm that they will respect it. This consent can be stored in the AAI components and further used as authorization data or forwarded to the service.
An exciting topic discussed further in this thesis is the digital identity in the environment of LS communities. As the platform might be provided internationally, implementing a complete digital identity

2. Authentication and Authorisation Infrastructure

framework is not optimal, primarily due to the need to verify the person's real-world identity and the correct association with the digital one. Setting up centralized points for verification is not acceptable due to the span across multiple countries. In each of them, the aspects of establishing such a point might differ and can result in reasons standing against this concept. Therefore, the environment has to rely on existing user identities available from external sources and ensure the identity has got a high level of trust instead.
An important topic is the set of legislative requirements that need to be fulfilled. Each country the environment spreads across might have a different collection of restrictions on the user identities enforced by the law. A crucial thing is informing users about the processing of their data. For example, in the European Union, a necessity is to comply with the GDPR[3] [8]. The country's local regulations, such as the Law about Cyber Security in the Czech Republic [9] or the Slovak Republic [10], might need to be considered.
In the following chapter, we list several types of digital identities available to be used as external user representations. We set up a set of criteria and evaluate different types of identities the users of such systems might already own. Based on the evaluation, we choose the best possible options for the needs of the selected domain.

3. General Data Protection Regulation

2 Digital identity types taxonomy

In the previous section, we have stated that the context where the digital identity resides influences the amount of attributes, their syntax, and semantics. Due to this fact, there are several types of digital identities the users might already have. In the following text, we present different sorts of them and analyze them. We also discuss the processes of establishment and applicability in the chosen domain. For each of the specified type, we provide a brief description, discuss its strong and weak claims, and evaluate the following properties:

- general level of adoption for the particular type,
- information value - the number of attributes and their quality,
- data freshness,
- the difficulty of the process to obtain the identity (financial and time aspects),
- an effort for machine processing,
- the level of trust,
- the possibility of forging the link between the real and digital world,
- applicability in the model environment.

We define five possible values for each of the chosen topics - *VERY LOW*, *LOW*, *MEDIUM*, *HIGH*, and *VERY HIGH*. In some cases, the evaluation might contain more than one value. The *VERY LOW* mark represents the lowest level. In case of difficulty or effort, it corresponds to the process being easy or requiring minimal activity. On the other end of the scale lies the value *VERY HIGH*. It refers to the process difficulty being very hard or nearly impossible. For each type, we evaluate these criteria in a table. The first column contains the particular measure, and the second column displays our evaluation. The third column contains the number of points we give the type for matching the desired level of a particular property. The amount of points is in

the range of one to five. Maximum points mark total accordance with the requirement, one almost none. Table 2.1 summarizes the ideal levels of properties the particular identity type should achieve.

Table 2.1: Requirements set on the digital identity

Criterion	Evaluation
Level of adoption	HIGH, VERY HIGH
Information value	VERY HIGH
Data freshness	VERY HIGH
Acquiring process effort	VERY LOW, LOW, MEDIUM
Machine processing effort	VERY LOW, LOW
Level of trust	VERY HIGH, HIGH
Possibility to forge	VERY LOW, LOW
Applicability	HIGH, VERY HIGH

The level of applicability is dependent on the score, which we count by summing all the received evaluation points, with a maximum being thirty-five. We set the following ranges for this property:

- **0 to 16 points** - VERY LOW,
- **17 to 21 points** - LOW,
- **22 to 26 points** - MEDIUM,
- **27 to 31 points** - HIGH,
- **32 to 35 points** - VERY HIGH.

2.1 Self-asserted digital identity

The easiest way of obtaining digital identity is when the users create it on their own. We name this type as a self-asserted identity. Establishing it requires the user to fill in a set of information requested by the identity provider, usually via a web form. A syntax validation

might be performed to verify that the received information satisfies the format constraints. Also, usually, no other validation is performed, especially not verifying the subject's real identity. The environment has to rely on a belief that the information provided is correct and the link to the real-world person is valid.

Such an identity type is probably the most common one. It is often used in social networks and basically in any services that need to identify users but do not care about their true identity. Implementation is not tricky. Obtaining this type of identity is usually not problematic. As already mentioned, the process might be a simple act of filling a web form. The person who wants to create the identity fills the required data and submits the application. Traditionally the data input consists of several fields that ask the user to fill in an identifier, name, email, and a secret password. In such a case, the difficulty of the acquiring process depends on the complexity of the form. A significant factor is also the user-friendliness of the environment where the user inputs the required information. The last vital property we want to point out is the easy processing of user's data. The processing happens automatically when the user applies, together with the possible data format validation procedure.

The first issue of this identity type is the data freshness. It depends on regular updates of the provided information by the user or how often the system in which the identity resides forces the user to do so. Another problem with this approach is potential weak data correctness. As no further verification of the data accuracy happens, anyone can fake it. Therefore impersonating somebody else is almost trivial. In some cases, if the registration requires the user to input an email address, an act of validating it might be performed. Unfortunately, this just filters out registrations submitted by bots or web crawlers. The amount of data is another property we could consider as a negative. Attributes content and their number depend on how much information is requested during the identity establishment act. The important thing is what subset of the requested information is mandatory. For the optional things, we have to assume they might not be available.

There exists an additional way of how the trust level of the identity can be increased. Many environments relying on such an identity offer an ability to register additional elements that can be later required in the authentication process. An example is requesting a unique code sent

Table 2.2: Evaluation of Self-asserted digital identity

Criterion	Evaluation	Score
Level of adoption	VERY HIGH	5
Information value	VERY LOW | LOW	1
Data freshness	VERY LOW | LOW	1
Acquiring process effort	VERY LOW | LOW	5
Machine processing effort	LOW	5
Level of trust	LOW	1
Possibility to forge	HIGH	1
Applicability	LOW	19

to a specific mobile device owned by the user. Next, the user would have to input the received code together with the authentication credentials. Apart from the physical device or token, users can use a biometric-based factor. This mechanism, however, does not increase trust from our point of view. Users pick additional authentication credentials by themselves. Therefore, the single ownership of such token and association with the real-world identity has not been verified.
A particular subtype we would like to point out are social identities. The persons usually create it by themselves. Apart from the things mentioned so far, the vital thing the identity has to fulfill is a reputation. For example, we can take a look at the pictures advertised by the subject holding the identity. We could also review the history of postings and published data. Thus, a reputation factor is vital to this type of identity. However, processing it in an automatized way can be difficult. Therefore, we consider this type of identity as a standalone source unusable for the modeled environment. The table 2.2 describes our evaluation.
We consider this identity as commonly used and suitable for automated machine processing. The process of obtaining an identity is usually straightforward. For the data freshness, it is challenging to decide what value we should evaluate it to, as it depends on user behavior. From a pessimistic view, we might assume that data is not updated regularly. Other things we are considering are marked as not

suitable for our needs. The information value can be relatively high, but because it is not further verified (corresponding with the trust level evaluation), we mark it as unsuitable for the environment of Life Science communities. Identity can be easily forged, as almost anyone can impersonate whomever they want. The overall applicability in the modeled domain is considered as low.

2.2 Self-asserted identity with verified secondary factor

Another type of identity that the user can create on their own is a self-asserted identity with verified secondary factor. It is similar to the self-asserted identity with the use of additional authentication factors described in section 2.1. In the mentioned similar type, the use of multiple authentication factors was optional. In this case, it is considered a mandatory part of the identity. The main difference lies in the level of trust of the user and the second-factor association. In the previous scenario, no proof of unique ownership of the additional registered factor exists. In the case of a step-up self-asserted identity, a third party establishes the secondary association and proves it is valid. An example of such a secondary factor can be a mobile phone provided by the employer.
As in the self-asserted identity, the additional factor does not have to be only a physical device. The user could use biometric information instead, e.g., fingerprint scan verified by a third party. When the user authenticates, the system requests usage of the second factor. In this case, the environment must believe that no other person has access to that secondary factor, as the third party has proved the association validity. If the user does not own such a token or cannot use it, using a work email address might be sufficient. Via verification process based on unique information sent to the specified email user can prove access to it.
This type of identity might be quite common, as many services have enabled registering the additional factor to increase their security aspects. Establishing such an identity might not be difficult. The critical part is the additional factor granter's policy if the user can use the given token for such purposes. The critical thing that needs to be solved is

Table 2.3: Evaluation of Self-asserted step-up digital identity

Criterion	Evaluation	Score
Level of adoption	MEDIUM	3
Information value	LOW	2
Data freshness	LOW	2
Acquiring process effort	LOW	4
Machine processing effort	HIGH	2
Level of trust	MEDIUM	3
Possibility to forge	LOW	4
Applicability	LOW	20

getting information about the token. The difficulty of the obtaining process depends on the difficulty of self-registration into the target environment. Except for a place to fill in user information, it is also necessary to provide tools for registering the additional authentication factor and checking its source. Usually, information about the token granter cannot be obtained in an automated way. It could require contacting the token-providing entity and manual verification of the token ownership claim or designing an automated procedure. The verified secondary factor association increases trust. It also lowers the possibility of forging such an identity, as the additional factor would need to be stolen or compromised. Information value is dependent on how much data the user will provide when registering and the mandatory dataset enforced by the system. Data freshness depends on how often users update it by their decision or are forced to do so by the system. In many scenarios, this need is put away in favor of user comfort. Therefore we take a pessimistic view and consider it as not well supported. Our evaluation is presented in the table 2.3.

As we can read from the table 2.3, this type of identity might be commonly used and easy to be obtained. The tricky part is getting information on the additional token source. Data freshness is another weak point. As the data is self-filled by the user, updating the attributes might be delayed or not happen at all. In the case of verification of the secondary factor claim, the trust level can get relatively high. This fact

also corresponds with the lower ability to forge the identity. In the information value aspect, we make a pessimistic opinion and consider it as not high enough. In total, this particular digital identity type does not seem to be applicable in the chosen domain.

2.3 Identity federations

Imagine a situation where several similar institutions provide their employees with digital identities and manage them. If the companies have common interests, it might be reasonable to use a similar identity management approach. Besides, it also makes sense if the users of one institution could use their original identity in the context of other subjects. The organizations might then create a grouping, which is called Identity Federation. The fact that the various providers have formed an association between themselves means that they must have a certain level of trust between themselves, sufficient to be willing to exchange messages with each other. When these messages contain the authentication and authorization credentials of users, allowing users from one system to access resources in a federated system, we have federated identity management (FIM) [11]. Applying this approach saves a lot of time and effort. Instead of solving the situation multiple times for its users and hosts, the organization creates a single point, where it is solved once.
The party managing the user account, performing the authentication, and possibly providing data about the subject to end service is usually identified as IdP[1]. The entity that consumes the identity attributes is referred to as a SP[2]. Instead of managing the user account itself, the SP delegates this role to the IdP. When a user wants to use the SP, they are forwarded to an IdP to perform the authentication. After successful completion, the user is sent back to the SP, which receives requested user information and can execute authorization decisions.
This model's plus side is that the SP does not care about managing the identity itself and not even about the authentication process. Such a model is quite used in some fields. The trust in the identity could be relatively high as the organizations usually establish the user's

1. Identity Provider - https://en.wikipedia.org/wiki/Identity_provider
2. Service Provider - https://en.wikipedia.org/wiki/Service_provider

digital identity via a formal process that includes verification of the real-world identity. Automated processing of the identity data is pretty simple from the consumer's perspective. Protocols created for federated identity management kept the idea of automated data processing. Organizations usually keep a comprehensive list of user data, and this information is required to be updated as soon as the real-world changes happen. The negative side of such a model is the possibility of limiting the set of users. To use a system provided in the federated environment, users must establish identity at one of the member Identity Providers. Also, the members need to be willing to provide authentication to the target services. The threat of an identity being forged depends on how well the Identity Provider has implemented the credentials protection. Furthermore, organizations can become a target for an attack on the identity management infrastructure more often. In the following sections, we evaluate the different federation's identities according to the before set criteria.
Service providers might form federations as well. These aggregations can provide users an infrastructure of services that can be used with identity established within the original IdP. Identity federations are formed in many areas. We want to point out the three most common types:

- academic federations
- identity federations formed by the government
- commercial subject identity provider federations

2.3.1 Academic identity federations

Federations of this type consist of the subjects operating in the education field. Members are usually universities, academic institutes, or other educational entities. These types of IdP federations are often quite dynamic. Subjects join them or leave them quite often. When a new subject fulfills the requirements, it can join the federation without much of a problem. An example of such a federation can be the SIR[3], an identity hub for Spanish academic and research institutions.

3. The RedIRIS Identity Federation - https://www.rediris.es/sir/

Table 2.4: Evaluation of the digital identity provided by academic identity providers

Criterion	Evaluation	Score
Level of adoption	HIGH	4
Information value	VERY HIGH	5
Data freshness	HIGH	4
Acquiring process effort	LOW	4
Machine processing effort	VERY LOW	5
Level of trust	VERY HIGH	5
Possibility to forge	VERY LOW	5
Applicability	HIGH	32

An interesting example in this field is the eduGAIN[4] inter-federation. It is an international service connecting research communities and higher education identity federations around the world, providing a single integration. [12] As it includes academic federations across the world, it is obvious that there might be some differences among them. To easily distinguish the quality of the identity provider, many frameworks like CoCo[5], SIRTFI[6] or R&S[7] have been introduced. These properties indicate that the identity provider satisfies the requirements set by these frameworks. In the table 2.4 we present evaluation of this type of the identity.

As presented in our evaluation (table 2.4, this type of identity is well established. Many academic institutions have joined their effort and created identity federations. However, this approach limits the set of possible users only to the people affiliated with an academic unit. Many of the users in the LS communities do meet this criterion. However, some of them do not. Therefore, we evaluate the level of

4. eduGAIN - https://edugain.org/
5. GEANT Data Protection Code of Conduct - https://wiki.refeds.org/display/CODE/Code+of+Conduct+ver+2.0+project
6. The Security Incident Response Trust Framework for Federated Identity - https://refeds.org/sirtfi
7. REFEDS Research and Scholarship Entity Category - https://refeds.org/category/research-and-scholarship

adoption to the value *HIGH*. Members of these federations usually store a decent set of precise and fresh information. The acquiring process often includes a formal procedure. It makes the process more challenging but in an acceptable amount. From the point of an identity consumer, the process is a matter of establishing a Service Provider for the service. Machine processing is suitable, and to forge such an identity becomes problematic as the legal process takes place. The identity trust level is satisfactory. Overall, this identity type might be the right choice for LS community infrastructures.

2.3.2 Government established identity federations

In the government area, the identity federations consist of government-approved subjects allowed to provide digital identity. Such an identity is primarily created and used in the e-government. Members of the federation (IdPs) usually have to meet requirements set by the law. In contrast to the academic field, not anyone can join the federation, only the designated and approved subjects.
An example of such a federation could be taken from the Czech Republic. Several identity providers like MojeID[8], NIA ID[9], BankID[10] and others offer this type of digital identity. They were approved by the government to do so and integrated into the electronic government systems. Table 2.5 summarizes our evaluation.
This federation type provides a good set of information. Data is fresh and suitable for further processing in an automated way. Trust level is high, and the possibility to forge such identity is very low. This situation is a result of the government playing a role in the federation and its establishment. However, the process of obtaining the identity might become quite complicated. As we are focusing on the international environment, we also have to consider differences between the countries. So far, no ultimate solution has been created and adopted by most countries in the world. There are significant differences in what the countries have adopted so far. However, if we look at a smaller set of countries, e.g., the European Union, the eIDAS regulation creates a base for an ultimate solution for these countries. The two mentioned

8. MojeID - https://www.mojeid.cz/en/
9. NIA ID - https://info.eidentita.cz/ups/
10. BankID - https://www.bankid.cz/en/

Table 2.5: Evaluation of the digital identity provided by government approved identity providers

Criterion	Evaluation	Score
Level of adoption	LOW	2
Information value	VERY HIGH	5
Data freshness	VERY HIGH	5
Acquiring process effort	MEDIUM	3
Machine processing effort	VERY LOW	5
Level of trust	VERY HIGH	5
Possibility to forge	VERY LOW	5
Applicability	HIGH	30

criteria force us (in the current situation) to evaluate as applicable but not as the ultimate solution.

2.3.3 Federations consisting of commercial subjects

Some companies with common interests have decided to act as IdPs in the commercial sector and formed identity federations. Such groupings can be quite dynamic, which is similar to the academic sphere. New members might join almost as they wish if they meet the requirements of becoming the participant. An example can be the Czech BankID[11] or the Finnish Bank ID[12]. It allows the use of bank identity to authenticate in the e-government or e-commerce systems. In the table 2.6 we evaluate this type of identity according to the needs of the Life Science community environment.

So far, this model has been adopted in some countries, but not on the cross-border level. We can often observe significant differences between the countries' approaches that have already adopted this approach. Information value is dependent on the set of collected information, which is usually reasonably sufficient for the needs of the chosen environment. Data is usually fresh, suitable for further automated machine processing. Forging such an identity is complicated

11. https://www.bankid.cz/en/
12. https://www.nets.eu/developer/e-ident/eids/Pages/BankIDFI.aspx

Table 2.6: Evaluation of the digital identity provided by commercial identity providers

Criterion	Evaluation	Score
Level of adoption	LOW | MEDIUM	2
Information value	HIGH	4
Data freshness	VERY HIGH	5
Acquiring process effort	MEDIUM | HIGH	2
Machine processing effort	VERY LOW	5
Level of trust	HIGH	4
Possibility to forge	VERY LOW	5
Applicability	HIGH	27

due to legal processes taking part in getting the identity. The establishment is usually not difficult as it commonly happens at some point in a person's life (e.g., getting a bank account). Overall this type of federated identity might be applicable in the LS community environment at a reasonable level.

2.3.4 Summary

As we can see from the tables 2.4, 2.5 and 2.6, we consider this type of identity as reasonably available, especially the education field. Most of the LS community infrastructure's users have a relationship with an educational or similar entity that might be a federation member. Information value can be relatively high but might also be limited by the IdP policy of releasing user data. Data is usually up to date and corresponding with the real-world situation. Obtaining a federated digital identity requires the user to become a member of the organization. It usually uses an administrational process, which happens to increase our trust in this identity type. The automated machine processing of identity data is a common goal in these cases. From the beginning, protocols developed for this purpose have kept this aspect in design. The ability to forge such an identity gets challenging as the formal verification process takes place. Overall, this type of user

representation might be one of the choices to take in the modeled environment.

2.4 Identity obtained via provisioning

Another type of identity that is created and maintained by a third party is a provisioned one. As mentioned in the case of identity federations 2.3, the identity establishment and management activities are delegated to an external entity. User attributes are then forwarded to the target consumer via provisioning. From the perspective of digital identity, provisioning is creating the identity record and its population with the correct attributes. [2] The difference between the two mentioned mechanisms is in the delegation of processes. In a federated environment, the authentication is performed by a third party. In this case, the authenticity verification is done by the target entity. Authorization stays to be the responsibility of the identity-consuming subject. The mechanism of provisioning comes together with the reverse process of de-provisioning. The result of this action is erasing the data when it is deleted in the source system. Failure to properly de-provision an identity can lead to confusion, access to critical data by outsiders, and even fraud or theft. [2]
An example of such an approach could be storing user accounts in a local LDAP[13] server. The system would rely on an external entity (e.g., a university information system) to handle identity management processes. When a user record would be created in the source system, provisioning would reflect this situation in the target system via creating a new entry for the identity in the LDAP directory. Similarly, the update process follows such a mechanism. The process of de-provisioning the identity might be considered optional. In case that the target environment would still make it possible for the person to use the systems even after the identity would get removed at the university, the entry in the LDAP could still be kept.
The process described is pretty straightforward. However, it relies on several properties of the connection between the source and destination. One problem might occur if the source does not perform the provisioning when the data is created or updated. In such a situation,

13. Lightweight Directory Access Protocol - https://en.wikipedia.org/wiki/Lightweight_Directory_Access_Pr

the destination stays uninformed about this fact. Also, the data can get corrupted during the transfer. The source needs to be informed about this event and repeat the process. The last thing we would like to bring up is the need for the provisioning mechanism to fulfill the confidentiality requirement.
From the favorable properties of this model, the target system does not have to provide any place where the user can obtain the identity. However, it needs to integrate with the provisioning mechanism. If we consider only the propagation and leave out the user interaction, data is usually quite actual. It is updated as soon as the provisioning mechanism propagates it. In an optimistic case scenario, this happens almost immediately after the change in the source. The ability to forge the propagated data depends on two things. The first is how well can the source system protect the data. Secondly, we need to ensure the information source is the correct one.
One of the negatives is that this model is not very used so far. The source and the destination have to agree on the data exchange. They need to specify the whole propagation process, including data exchanged, mechanism of data forwarding, and security aspects. Information provided by the source can be limited due to regulations the provider needs to fulfill. Establishing the identity is dependent on the difficulty of this process at the source. It can vary from being easy to difficult. The process also affects the level of trust we can put into the provisioned data. In the table 2.7 we provide our evaluation of the model.
In summary, this model is not widely used yet. It provides sufficient but not extensive information value. Data is usually fresh, but this property depends on how often the source enforces data updates. Obtaining the identity might not be trivial nor complicated. Data is usually suitable for automatized machine processing. The level of trust can be relatively high. It depends on how well the source verifies the identity attributes. The ability to forge the association depends on several things - the implementation of the authentication mechanism, how well the source protects data against the attacks, and the security of the provisioning mechanism.
A problematic side of this model could be the size of the modeled domain. We are focusing on the international environment. Due to this fact, we would require an internationally used system to provide us

Table 2.7: Evaluation of the digital identity provided via provisioning

Criterion	Evaluation	Score
Level of adoption	LOW	2
Information value	MEDIUM	3
Data freshness	MEDIUM	3
Acquiring process effort	MEDIUM	3
Machine processing effort	VERY LOW	5
Level of trust	MEDIUM	3
Possibility to forge	MEDIUM	3
Applicability	MEDIUM	22

with such capabilities or integrate multiple systems, e.g., per country. When implementing such a solution, several components would serve as middleware for collecting the data, unifying the format, and similar activities. The overall scalability of this model is not at the level we would require. In total, this type of identity is not too applicable to the modeled environment.

2.5 Reputation-based identity

As we have already mentioned in section 2.1, reputation can be an imporant part of digital identity. The identity has to gain a reputation at the required level to gain trust by other participating entities. Subjects can evaluate the trustfulness of the actions taken and the overall behavior of the identified subject. Another thing we can observe is the relationships between subjects.
An example of this concept is the PGP[14]. It is a standardized encryption system providing means for performing cryptographic operations, primarily encryption and digital signature. To enable the concept, users have to create a pair of credentials - public and private key. These keys are then used when performing the cryptographic actions. In the context of identity, we can assume that any data on which a cryptographic operation has been performed using these credentials is related to that

14. Pretty Good Privacy

user. The keys are unique for each user and thus identify them. An example can be the digital signature of a data set.
An important thing we have to mention is the cross-signing of the keys. It is the base concept upon which the PGP trust architecture is built. Instead of a hierarchical trust architecture with Certification Authorities as in X.509[15], OpenPGP employs a certification model where any entity can certify another entity, which results in a so-called Web of Trust (WoT). [13] The reputation rises with each signature of the public key made by another user. The problem is that in the PGP, the reputation property never decreases. The only way of downgrading the reputation is the act of credentials revocation. However, this action invalidates the keys, and we cannot state that until some point in time, e.g., the digital signatures, are correct and, from that point on, are incorrect. The whole dataset is treated as not signed at all.
An identity can be build based on a similar mechanism as we have described above. Let us assume that each user of the target system has got such keys. For example, to establish the identity, the user would need to sign a piece of information using their private key. By using the public part of the pair, the signature correctness could be verified. Such a mechanism could also be used for authentication. To increase the association's trust with the keys, the LS community could generate the cryptographic keys and grant them to the users. Another option would be to use the already existing keys and verify a real-world person's identity.
Another reputation-based approach is the method of identity vouching. It is similar to the web of trust concept but not based on the keys. In such an environment, one or more different subjects state that the link between digital representation and the real person's identity is correct and can be trusted. This mechanism's problem is the trust in the subjects making the decision and possible cooperation of evil subjects. Such a group of attackers could state that a lot of correct associations are invalid and vice versa.
If we look more at this concept's properties, the level of adoption can be relatively high. Cryptographic operations and remarkably digital signatures are commonly used nowadays. Establishing an identity

15. Standard defining the format of public key certificates - https://en.wikipedia.org/wiki/X.509

framework based on such an approach is non-trivial but not impossible. The information value can achieve potentially high levels. All the operations performed on the data are made digitally. Therefore, machine processing of the data is significantly easy. As we have mentioned, the level of adoption of cryptographic keys is relatively high, and it has been reflected in how the credentials pair is generated. Therefore, the process of obtaining an identity might be comfortable as well. Generating cryptographic credentials is not tricky. The only hard part of the process could be linking the real-world identity to the user. On the pessimistic view, keeping the data about the identity owner up-to-date could become problematic. After creating the credentials, users might not update the associated descriptive information too often. Also, security issues need to be considered, as the risk of compromised credentials might be high. Gaining access to some person's private key is nowadays becoming a reasonably valuable thing for the attackers.
An important thing is a need for users to be educated on how to use the credentials. Another thing that needs to be mentioned is the possibility of stealing the credentials. Users have to be instructed on protecting it and what to do in case of credential leakage. As the keys are digital information, the risk of unauthorized access is relatively high and does not have to be noticed immediately. It might even enforce some physical protection, which is still not the ultimate solution to the problem. Therefore, the possibility to forge such an identity can rise to unacceptable levels as well. Our evaluation is summarized in the table 2.8.

The base mechanism for this identity type is widely used. Information value is dependent on how much descriptive data is associated with the credentials. Usually, some necessary data needs to be provided when generating the keys, and this set might be sufficient. However, it needs to be further validated if the data is correct. Data freshness might be at a low level. Revocation of the credentials might cause potential issues and needs to be correctly handled. The process of obtaining the credentials (cryptographic keys) is relatively easy. The only tricky part might be associating it with a real-world person. The ability of the identity to be processed by a machine is convenient. The level of trust and possibility to forge is not at the levels we would require. In total, concerning the possible problems with such a sys-

Table 2.8: Evaluation of the digital identity provided via provisioning

Criterion	Evaluation	Score
Level of adoption	HIGH	4
Information value	MEDIUM	3
Data freshness	LOW	2
Acquiring process effort	LOW	4
Machine processing effort	VERY LOW	5
Level of trust	LOW	2
Possibility to forge	HIGH	2
Applicability	MEDIUM	22

tem's identity and scalability, we evaluate it as not too applicable in the context of Life Science community infrastructure.

2.6 Physical document based digital identity

Digital identity can also be based on a physical identification card. We are talking about a passport, id card, or a driving license in an ideal case. However, other physical documents can be used as well, e.g., a student identification card might be sufficient. We want to point out that we are not interested primarily in the computational abilities of the card in this type of identity. Our concern is establishing an identity that is, e.g., self-asserted, and the document is used for verification via an automated process.
An example in this field might be a mobile application, in which the users fill out a set of information about themselves. After that, the application would require them to scan their identification card or passport using a camera or NFC[16] technology. A machine would then perform automated cross-verification of the data parsed from the scan and user input. Another possibility would be remote manual cross-check by a designated employee of the application operating company. The process might incorporate validation of the person's liveness by a

16. Near-field Communication - https://en.wikipedia.org/wiki/Near-field_communication

short video call or requirement of recording a video at the moment of registering the identity.
An interesting topic in such a case is the trust of the person in the application. As it requires scanning the document, which in most cases contains sensitive data, the belief in the scan not being misused is crucial. Another essential thing is detecting the provided document is authentic, and not just an already scanned one. It can be solved using liveness detection by requesting a recording of a video scan of the user's face.
If we look at the properties of such digital identity, the outstanding is the identity's ability to be processed automatically. Parsing the identification data from document scans is nowadays a commonly used approach. With the possibility to use NFC for scanning the digital version of the data, this process can be even more straightforward. Looking more closely at the adoption level, we have to state that it can be potentially high. The main reason is that the process of obtaining the identity is not too complicated. Some of the events that might stand against its further adoption are the problems we have described in the previous paragraph. Due to the requirement of using some formal document, the possibility to forge the identity decreases. It is the result of having to forge the document. Detecting the falsification of the document is crucial. The formal document also influences the level of trust. As expected, it increases with the level of trust of the used document. However, exacting such information might be nontrivial. From the weak attributes, we have to mention data freshness. Users might establish the identity, and after the verification, the process stops. The data might not get updated in the future too often. Information value can be a weak property as well. The evaluation would depend on the type of physical document used for verification, what data we receive from it, and what data stays unverified. Table 2.9. summarizes our evaluation.
As we can read from the table 2.9, this type of identity might become well adopted. The information value might be low, but we can assume receiving more than just necessary information like the person's name. A valuable thing to know would be what kind of document has been used for establishment. Data might not get updated very often. The process to obtain the identity can be relatively easy and user-friendly. The verification part might be automated, and therefore

Table 2.9: Evaluation of the digital identity based on physical identification document

Criterion	Evaluation	Score
Level of adoption	MEDIUM	3
Information value	MEDIUM	3
Data freshness	LOW	2
Acquiring process effort	LOW	4
Machine processing effort	VERY LOW	5
Level of trust	HIGH	4
Possibility to forge	MEDIUM	3
Applicability	MEDIUM	24

the machine processing effort is marked as suitable. Trust level rises with the improved verification process and more formal documents used for it. As the procedure includes scanning the document, the possibility to forge the association gets more challenging. The main problem we would have to face is convincing the users that their document scans would not get misused in the future. In total, this approach to the identification of users is not suitable as the primary source of identities.

2.7 Public Key Infrastructure

As mentioned in section 2.5, cryptographic keys can be treated as a digital identity. A digital certificate, or often referred to as a public key certificate, is a document confirming ownership of the public part of the cryptographic keys. Digital certificates are at the heart of protecting all aspects of data communication, from websites for business and banking to shopping and product development to social media for interaction and collaboration. [14]
The identity model as a digital certificate is based on the ownership of a pair of credentials for asymmetric cryptography. Using the private and public keys, the user can perform cryptographic operations, such as creating digital signatures. If we, besides, have the certificate of

ownership, we have a piece of verified information, who is the owner of the credentials and therefore can identify the subject.
If we look at the certificate's content, it contains the public key and identification of the owning person. The owner part can contain fields like the name of the subject, country, organization, or alternative names. As well as the data describing the subject and issuer, the certificate contains information about the issued document's validity time. After the period ends, the certificate is considered inactive and has to be replaced by a new one. It is a vital concept that can be used for the requirement of data freshness.
In recent years, much effort has been made in adopting digital certificates as an ultimate digital identity solution. A technology abbreviated as PKI[17] has been developed. It is a method of authenticating users and devices in the digital world using asymmetric cryptography mechanisms. PKI environment consists of several entities. The basic one is the user (or machine) holding the keys. Then, we distinguish the trusted parties - CAs[18]. They validate the credentials of the owner and sign the digital certificate. The CA's signature validates that the owner is who they say they are and that the public key belongs to them. The CA might also generate the key pair if the person or server has not already done so. [14] Authorities can form hierarchies if needed. Most infrastructures have adopted the X.509 standard defining the format of the certificates.
If we take a closer look at such identity's properties, the adoption level is becoming higher, as digital signatures are getting used more and more. Information value depends a lot on the identification information included in the certificate or published by the CA. Dataset describing or identifying the person can be quite extensive. The provided information is not ultimately fresh but has to be renewed when the certificate expires. The machine processing is suitable, as the identity does not need to be further verified. The level of trust is another from this type's pluses, as a trusted third party creates the association. The possibility to forge it is not critical as well. The problem could occur if an unauthorized person gets access to the secret key. Many users own the digital certificate but do not necessarily use it for au-

17. Public Key Infrastructure - https://en.wikipedia.org/wiki/Public_key_infrastructure
18. Certificate Authority - https://en.wikipedia.org/wiki/Certificate_authority

Table 2.10: Evaluation of the digital identity based on physical identification document

Criterion	Evaluation	Score
Level of adoption	VERY HIGH	5
Information value	HIGH	4
Data freshness	MEDIUM | HIGH	3
Acquiring process effort	LOW | MEDIUM	3
Machine processing effort	VERY LOW	5
Level of trust	HIGH	4
Possibility to forge	LOW	4
Applicability	HIGH	28

thentication. The effort the user has to make to get a digital certificate is nontrivial.
In some cases, the verification process might include confirmation in person. In other, the certificate is issued with the belief that some external organization provides the data about the user (e.g., university). However, it might include a formal procedure of the user identity verification preceded with an application for the certificate, filling in information about the subject, or other.
We want to point out the need for users to be educated on using and protecting the credentials. As the trend of using digital signatures rises, gaining access to somebody's private key without the person noticing it might cause serious damage. It has triggered enormous development of mechanisms for protecting the private key by using different passwords, biometric protection, or storing the private key on an external hardware component. Table 2.10 contains our evaluation of type of digital identity.
As we can conclude from the evaluation, we consider this identity type's base concept widely adopted. It can provide enough information about the identified subject. Data is refreshed at periodical intervals, at least when the certificate expires and needs to be renewed. The process of obtaining is nontrivial but at an acceptable level. Digital certificates are suitable for automated machine processing. The level of trust is quite significant. The possibility to forge the existing identity relies

on how well the subject can protect the credential's private part. The overall risk is acceptable, as further development on protecting the private key continues. In summary, this type of identity corresponds to our requirements at a reasonable level.

2.8 Digital identity established by a commercial subject

The commercial sphere is another place where to look for digital identities. People often apply for some discounts, buy a product or a service. Especially in online shopping, they are provided with a user account representing them and providing access to manage their preferences or communication with the business subject. In some cases, validation of the real-world person's identity needs to happen. An example of this could be registering for a user account in a health insurance company. With such a digital identity, the person could then access the company's systems to make transactions online. As the business needs to ensure the identity of the person represented by the account, robust verification of the identity, e.g., by providing a scan of the identification cards, needs to happen. Often, such an account can be based on an already issued document or before made contract.
As a result of such processes, the person often establishes a strongly verified digital identity managed by the commercial subject. National institutes might even enforce publishing identifiers for these identities in standard registries. A set of identifiers for a single person could be created from multiple sources, with the vision that these identifiers could be cross-linked together. An example of such a set could be identifiers from the health insurance company, bank, mobile phone operator, internet provider, or public utility providers. Such a collection of identifiers could be used when the authentication process happens. As in the idea of federated identity 2.3.1, users could use the digital identity provided by a commercial subject to get access to a desired digital environment.
If we take a closer look at the positives of such an approach, we need to point out the information amount these identities could provide. As subscribers of such subjects usually sign some contract with the provider, these entities require applicants to fill in a nontrivial amount

of personal information. Another strong side is the data freshness. Having up-to-date user information is crucial for these subjects. Users of such systems are usually obliged to update the information that changes as soon as it happens. Machine processing would be a low effort operation as the identity is usually strongly verified and identity is already digitalized. The level of trust in such an identity could be potentially high due to formal verification. Also, the possibility to forge such an identity is relatively low.
On the negative side of view, the acquiring process can get quite complicated due to its formal requirements and often nontrivial form filling and paperwork. The level of adoption of such an approach is not at the desired level yet, especially at the international level. A problem that might occur in utilizing this concept is the legal agreement granting subjects access to the joint registries. As a consumer, one could be denied accessing it, and the identifier provided by the commercial subject would become useless in the target system. A problematic situation might occur with linking multiple identifiers from multiple subjects. As it might be helpful to link them together, the law regulations might stand against doing this. So far, we have not reached the status where these identifiers could be aggregated into a set. In the table 2.11 we summarize our assumptions and evaluations of the properties for this type of digital identity.

Table 2.11: Evaluation of the digital identity established via commercial subject

Criterion	Evaluation	Score
Level of adoption	VERY LOW | LOW	1
Information value	MEDIUM | HIGH	3
Data freshness	VERY HIGH	5
Acquiring process effort	MEDIUM	3
Machine processing effort	LOW	4
Level of trust	HIGH	4
Possibility to forge	VERY LOW	5
Applicability	MEDIUM	25

Overall, we have evaluated the adoption level as very low due to the concept gaining some popularity but still not being used widely, especially not in the cross-border environment. The amount of data provided by the identity could be sufficient. Data is usually refreshed as soon as it changes. The acquiring process might become more complex with an increasing number of forms to fill and paperwork. Automated machine processing is at a suitable level. The level of trust could be relatively high. The possibility to forge such an identity is potentially low due to the strong verification process. In total, this type of identity might be applicable, but it is not an ideal solution.

2.9 Electronic identification

As more and more of the world moves towards the digital environment, so have done nation's governments. Identification cards went through considerable development. For example, many identification cards now have got a microchip and memory. The ID card can perform cryptographic operations and store different identification data, such as biometric information. In this chapter, we will consider passports and state-issued identification cards as a source of digital identity. We mainly focus on the capabilities of the cards.
An excellent example of development in this area is Estonia. Its inhabitants can use the identity cards issued by the government for creating qualified digital signatures. It can also be used to authenticate for an electronic vote, as a proof of identification for logging into the bank accounts, used as an electronic medical prescription, or as a national health insurance card. It contains the certificates that associate the document holder with the card's activities in the digital environment. These certificates are protected with PINs given to the owner in a sealed envelope upon the ID card issue. The last two components needed are an internet connection and specialized hardware to communicate with the embedded microchip. A specialized application has been developed from the software point. It allows the use of the ID card as a token. [15]
Activities to develop such a framework are getting more and more supported by the governments. For instance, in the European Union, the eIDAS regulation has enforced the use of digital signatures. If

we go through the document, we will find out that it enforces the digital signature to be evaluated at the same level as the signature of a document made by the person in the real world by hand. The electronic identification (eID) framework is supported by the EU a lot as well. It encourages governments to provide services in a digital environment, e.g., an electronic vote system. To do that, inhabitants of such countries have to authenticate into those systems with a verified identity. As a result, more and more countries have enriched personal identification cards with a microchip containing unique credentials for each person. By plugging the ID card into a particular hardware component, the subject can authenticate into such a system, e.g., by performing cryptographic operations.

From the positive attributes of this identity type, we have to mention the information value. National ID cards usually contain extensive sets of identification data. Often it also contains some biometric information. Data freshness is another strong point. The data is usually updated very soon after it changes due to the obligation of updating the identification document when some attribute changes (e.g., change of surname or residency). Machine processing is another point at which this type of identity excels. The level of trust is considered very high. It is a result of the identification card issued by the government. The security is increased by protecting the identification certificates with a secondary factor - PIN code. These facts also result in a very low possibility to forge this identity as it is hard to forge the document and gain access to the security codes. On the negative side of things, the level of adoption is not yet reaching the point we would like. Only some countries have developed the means for using such identity at a satisfactory level. Another important thing is adoption from the side of services that want to make use of it. They need to fulfill some set of rules to be able to use the identity. The acquiring process gets more difficult as an official procedure has to go through to obtain the ID card. Besides, the user has to own the required additional hardware and install software to use the identity. We summarize our evaluation in the table 2.12.

As we can see from the evaluation, the level of adoption is not satisfactory. Information value is considered very high due to the microchip presence on the smart card. It can contain a potentially large amount of identification data, which is usually very fresh. The effort to ob-

Table 2.12: Evaluation of the digital identity established via electronic identification

Criterion	Evaluation	Score
Level of adoption	MEDIUM	3
Information value	VERY HIGH	5
Data freshness	VERY HIGH	5
Acquiring process effort	MEDIUM	3
Machine processing effort	MEDIUM	3
Level of trust	VERY HIGH	5
Possibility to forge	VERY LOW	5
Applicability	VERY HIGH	29

tain the identity is higher as a formal process is needed. A machine can efficiently process user data as the whole concept has been built around this process. However, additional hardware and software requirements make it more difficult. The level of trust is very high due to the document being hard to forge and using a secondary authentication factor (PIN). The possibility to forge is very low due to restrictions on who can issue the document. In total, this type of identity is considered as well applicable in the infrastructures operated by Life Science communities.

2.10 Combining multiple identities

So far, we have considered the identification of the users by only one type of identity. However, we could combine multiple representations and therefore increase overall applicability in the modeled environment and user-friendliness. By combining various approaches, a new digital identity is created. This new instance can leverage information and trust from the linked ones. For example, we could combine identity with the highest possible trust level with a commonly used one. In such a situation, we would achieve a very high trust level and comfort of the user by applying the well-adopted identity type mechanisms. To give an example, we could combine a federated academic

identity with a self-asserted one. The federated identity provides us with enough information as well as data freshness and level of trust. Also, it cannot be forged easily. The self-asserted type has a strong side in the common usage and familiarity from the user side. For some properties, like the user's effort to obtain those identities, we need to consider both levels together, as the user needs to go through both processes.

In some cases, we need to consider all the property values from linked identities and pick the lowest value as the final value for such property. An example of such a case is the possibility of forging the identity. If one of the chosen identity types could be falsified easily, then the whole identity created by the linking could be misused by the attacker. We could solve this by restricting what type of identity needs to be used for authentication in specific situations or adopting a mechanism to increase the set of privileges of the authenticated subject (e.g., authentication at a certain point).

We want to mention that some intermediate identity provider or specialized software needs to handle such cases and provide the means to create such a combined identity. For example, it could create a user representation based on the initial external identity used. Then all the following identities of a user would be linked to this representation. The joint user representation would be the resulting digital identity. This approach has many benefits as user comfort when using the identity and extended resource set for attributes. Unfortunately, there are also some flaws. As mentioned in the previous paragraph, an example is a security risk due to possible more significant security issues of one of the linked identities. The identity linking process must also be well designed to guide the user through it and not confuse them.

2.11 Summary

Table 2.13: Comparision of mentioned digital identity types

Criterion	Acad. fed.[a]	Digital cert.[b]	E ID[c]
Level of adoption	HIGH	VERY HIGH	MEDIUM
Information value	VERY HIGH	HIGH	VERY HIGH
Data freshness	HIGH	MED.[d] | HIGH	VERY HIGH
Acquiring process	LOW	LOW | MED.	MEDIUM
Machine processing	VERY LOW	VERY LOW	MEDIUM
Level of trust	VERY HIGH	HIGH	VERY HIGH
Possibility to forge	VERY LOW	LOW	VERY LOW
Applicability	VERY HIGH	HIGH	VERY HIGH
Score	32	28	29

a. Academic federation
b. Digital certificate
c. Electronic identification
d. MEDIUM

In the previous text, we have discussed several digital identity types that the users might already own. From the evaluation performed for each class and then comparing the general results, we have picked the best possible options that suit our needs. In particular, we will further consider only identities from academic federations, digital certificates, and electronic identification. However, the other options are still applicable if the subject is already represented in the target system using some of the chosen approaches. The environment could allow linking multiple external identifications for a single user to a central representation in the community infrastructure.
As we can see from the table 2.11 comparing the most applicable solutions, these are the options that suit the needs of the target domain the most. They provide enough identification information about the user. Data is usually as fresh as possible. This fact is often a result of a formal agreement between the provider of the identity and the user

(work agreement, identity provided by the government). The process of acquiring the identity might get more complicated. However, this is a sacrifice we must make to obtain better properties in other aspects of our evaluation. The main reason for the process getting more complicated is the mentioned possibility of formal agreement or requirements set by the providing subject. Processing the identity attributes in an automated way is usually very easy as the chosen options were designed for such purposes. The level of trust is in all of the possibilities we have chosen at a reasonable level. Again, this results from the user and the identity provider's formal agreement created during the obtaining process. Selected digital identity types are not easy to forge due to the verification process that is usually performed. The important thing is educating the users on protecting the identity credentials or physical credentials.

3 Establishing digital identity in a distributed environment

As we have already stated in section 1.4, implementing the whole digital identity framework is not optimal in the distributed and international environment. It is common to rely on already existing user identities instead. As a result, the relying services do not need to think about how the identity will be established and often even managed. However, it still needs to set up the processes described in section 1.2. It still has to have its user representation and maintain the link with the external identity. The authentication process often changes to performing identification based on the identifier received from the external identity provider after the user authenticates there. The most important thing the service has to solve is the process of associating the foreign user identity with its internal representation.
The design and implementation of this process remain critical. It has to satisfy the needs of the service while providing a smooth user experience. The solution should be easy to adopt and based on already existing mechanisms. For instance, standardized protocols developed for these purposes can be used. Another vital thing is fulfilling the legal aspects that the solution needs to satisfy. Therefore, the processes should be documented and results made available to the interested subjects.
In section 1.4 describing this thesis's domain, we have introduced the concept of AAI. This framework is often adopted in an environment similar to the one of our interest. It can serve as a joint point of the services with their users in such an environment. The AAI can solve the necessary processes (identification, authentication, and authorization) both from the users and service perspective. As a result, the services can focus on the functionality provided rather than solving these usual procedures.
In the following sections, we are consulting the process of associating the external identity with representation in the community infrastructure environment from each of the perspectives. We describe the process's goal, what it should look like, and the requirements it has to

fulfill. We also evaluate a sample environment of our choice and its design and implementation of such a process.

3.1 User perspective

If we consider the goal of the user, it is to use a relying service. Let us imagine we are such a person. In an ideal case, we want to visit the service, perform authentication if needed, and use the functionality offered by the service. If authentication is needed, we must set up an identity representing us in the particular service. In the case of the environment offering multiple services, we should not create an account in each of them. Instead, we should be able to use a single identity across all of the services provided.
The first thing we might be interested in is information about the legal aspects and data processing. The documents established for these purposes need to be easily findable. Also, we want to navigate through different sources of such information as little as possible.
By setting up an account, the service will establish an association between the external identity and its user representation. This action is often realized by going through a registration process or via authentication with an external identity, while registration happens in the background. First, we need to discover the entry point to trigger the registration. This process should not require much effort. An easy solution might be registering directly at the service by using a designated webpage. An important thing to keep in mind is that we want to minimize the amount of information provided to the service. It should be explained why a particular set of information is required and how it will be processed.
From the assumptions of using external identity, we have to provide data about it to the registration system to create the association. As a third party manages the identity, its attributes might be hard to find or utterly unknown to us. Therefore, the process of fetching the identity attributes and especially the identifiers should be made transparent for us. A good example might be performing authentication with the foreign account and receiving the authentication information at a specific component. It should fetch and prefill the identity data into the registration form. As the identity might contain data that is not up to

date, we should be allowed to correct it. However, we should be permitted to edit only a subset of information, e.g., the email address. We should not be allowed to change data such as the external identifiers. The service might use several sources of foreign identities, like supporting a large federation of identity providers. Therefore, we need to be able to choose the correct external account we want to use. A component listing all the integrated identity providers might be needed. However, as the number of possible options might be nontrivial, the interface must support locating the correct entry representing our choice without difficulties.
When the registration consists of multiple steps, we should be navigated through all of them automatically by the registration component. Getting information about the current status of the process is crucial. In an ideal case, we should also be informed on the remaining and finished steps to keep an overview of the process. If an error occurs, we need to be notified about what has happened and how to solve it. The whole process should also let us leave and return at a different time to continue at the exit point.
After finishing the registration, we should be allowed to start using the particular service. Due to infrastructure storing data about us and possibly forwarding it to the service, we need to be able to review it. As time progresses, some of the information might get invalid and should be corrected. Therefore, an ability to update the infrastructure identity attributes must be available. We should again be allowed to edit only the data we understand from a semantics point of view. However, most of the data should be updated without our interaction, if possible.

3.2 Relying service perspective

From the perspective of the relying service, its primary goal is providing functionality to the users. It needs to identify the user, perform the authentication, and obtain the required data for performing the authorization decisions. Again, we want to remind, that we are assuming the service and the infrastructure which provides it (or integrates it) relies on an already existing user identity managed and provided by a third party.

In an optimal case, the authentication, and partially the identification processes, are delegated to the subject providing the identity. To some extent, the authorization can be done by the identity provider as well. After the user successfully authenticates, we, as a service, should receive attributes of the identity and information about the authentication. The vital part of received data is the identifier of the user identity. This information will be used by a service to identify the particular user in its context. Based on the received information, the service can perform authorization decisions. After all of the processes are finished with a successful result, the user can use the service.
The first step in the process is letting the users establish their representation in the service. In particular, we need to create it and associate it with the external identity. We can implement this process by requiring the user to fill a registration form. A user could be required to fill in the fields for the necessary information about themselves and the foreign user representation. As we rely on an external identity, it should be automated as much as possible. For example, we could allow users to authenticate at the identity provider and prefill the form with the received identity attributes. An even better solution would be to perform the process without user interaction, apart from the authentication.
Limiting the users to use only one external identity provider can result in low interest in our service. Instead, we should integrate with multiple identity providers. If supporting a number of them, an important thing is letting users choose the one they want to use. Such a feature can be implemented by providing a designated web page listing all of the possible options and redirecting the user to the external location based on the choice made.
As mentioned earlier, after the user authenticates, we should receive the identity attributes. The data from foreign sources might not be up to date. Due to this fact, we should allow users to review the information and possibly correct it. However, some fields should not be allowed to be edited by the users as they might contain critical information the user will probably not understand (identifiers and similar). Suppose we believe the attributes are up-to-date and contain all of the required information, or we just want to make the registration process more straightforward. We might skip the validation and modification steps performed by the user and return to it when the

identity is already established.
When the registration is submitted, a critical point is its verification. We need to ensure that the data is in the correct format and valid. The important thing is verifying that all of the required information has been obtained. We should then store the information and possibly process it to derive additional attributes. The critical part is associating the external identity identifier with the internal user representation.
When the user successfully establishes the identity, they should be able to start using our service. However, we need to recognize the particular user when they access it. In an optimal case, we could consider the user as already authenticated, especially if they have done so during the registration. After successful authentication, we should retrieve all necessary data for performing the identification and authorization. In the external authentication case, we should receive the identifier of the used digital identity from the third party and retrieve the data associated with this identifier from our internal storage. At this point, we can perform the authorization.
To increase user comfort, we might introduce a concept of account linking. In such a scenario, we should allow users to associate multiple external identities with a single user representation in our service. It could be beneficial for both the users and us. In our perspective, the additional identity attributes can extend the current set of information and increase the information value of the identity. A designated component allowing to perform the linking of the new identities should be provided in such a case. It can, for example, reuse the component for registration. Another essential thing to mention is letting the user also remove the association of external identity. Such a feature might be usable in the case of an external user account being compromised.
After some time of using our service, the information about the user can get outdated. It is crucial to keep the identity attributes fresh for performing correct authorization decisions. An example of ensuring data freshness could be periodically requesting the user to review and possibly update it. Another possibility is to implement an automated mechanism of refreshing the data. The first of the two approaches creates additional constraints on the validation of the modified data. The latter is an elegant solution when we rely on external digital identities. An example could be to refresh the data each time the user authenticates, and we receive the attributes from the external digital

identity.
Letting users display the information held about them is essential. Therefore, we need to provide users with a place to review and possibly correct the data associated with their identity in the service. Such a place might also act as an entry place for specifying different user-specific service settings, like localization, settings of email notifications, or similar things. As we further process user data, we need to set up policies and processes for managing it. This need can require us to create documents that should be (possibly in a different, more user-centric form) published and made available to the users as an information source on data processing.
As seen from the previous text, it is crucial to define the lifecycle of the user identity in the service. It should include all the stages of its existence in the service - the establishment, using and managing it, and decommission. This process can also be documented and published to the users to provide a clear view of their data processing.

3.3 Protocols for exchanging user identity information

One of the requirements we have set up is the ability to quickly adopt the chosen type of digital identity in the LS community infrastructures. In detail, we need to ensure that the process of delivering the external identity is secure and can be easily and quickly implemented. Therefore, the process should be based on standardized mechanisms that have been already developed and tested out. According to the selected types of digital identity we want mention the SAML 2.0[1] and OIDC 1.0[2] protocols. These standards are used for exchanging authentication and authorization information between participating subjects. Several other protocols exist and can be used as well. However, we consider the mentioned two as the currently most developed and used ones due to our experience. The following sections provide a general overview, concepts, and important roles of the entities taking part.

1. Security Assertion Markup Language 2.0
2. OpenID Connect 1.0

3.3.1 Security Assertion Markup Language 2.0

The SAML protocol is an XML[3]-based protocol used to transfer authentication and authorization information. It exists in several versions, with the latest being SAML V2.0, which we refer to further in this chapter. The protocol was developed by the Security Services Technical Committee of OASIS[4]. It allows the business entities to make assertions regarding the identity, attributes, and entitlements of a subject to other entities, such as a partner company or another enterprise application. [16]
The subject participating in the communication can act as an Identity Provider (IdP) or (and) Service Provider (SP). An entity operating the role of the IdP is responsible for managing the user identity and releasing data to the SP. The Service Provider is the entity consuming the user data. Each of the participants is described by a set of data referred to as metadata. It provides information about the entity and supported functionality. To communicate via the protocol, the parties have to first exchange the metadata with each other. The information set can be published together with other participating subject's metadata in a centralized way by joining a SAML-based federation. Establishing a federation enables users to utilize the concept of Single Sign-On[5] (SSO) across the Service providers consuming the identities from the federation.
An important thing this protocol makes use of is a concept of an assertion. It is a set of information conveying the statements supplied by the SAML authority. There are three basic types of these statements - authentication, attribute, and authorization assertions. The first type carries the information about the authenticated subject and authentication process. Attribute assertion contains the data about the user. The last type expresses allowance or denial of access to a specified resource.
Attributes of the digital identity are mapped to assertion attributes using a mapping schema. It is a set of rules specifying syntax, semantics, and identifiers of constructs used to transport user information. A significant amount of predefined attributes have been established

3. eXtensible Markup Language
4. Organization for the Advancement of Structured Information Standards
5. Single Sign-On - https://en.wikipedia.org/wiki/Single_sign-on

in different areas (e.g., the eduPerson [17] schema for the academic environment). Different schemas reflect the needs of the environment in which the protocol is used. Especially in the academic environment, the mentioned eduPerson [17], SCHAC [18] and the voPerson [19] schemas are important. Attributes defined in them describe the subject in the higher education area, like unique identifiers in the academic environment, affiliation to the home organization, principal names, name of the home organization, or similar. For example, the *eduPersonAssurance* attribute contains information about how identity verification process. The value obtained in this attribute can be used to express the level of trust in the identity.
An essential thing in the protocol is the ability of the user to control the data released to the SP. Before the attributes are sent, the user can decide if the transfer should be allowed or not. However, the subject cannot specify any subset of the attributes that should be released. The decision allows only to transfer all the requested attributes or none of them.
The primary environment in which the protocol has been developed was the academics. Many corporations have also adopted it after the success it has made. However, nowadays, many subjects prefer protocols that provide more straightforward implementation, metadata management, and additional features.

3.3.2 eIDAS SAML

The eIDAS regulation [5] ensures that people and businesses can use their national electronic identification schemes (eIDs) to access public services in other EU countries where eIDs are available. As a result, it creates a European-level trust network on electronic services (e.g., digital signatures) by ensuring that they will work across borders and have the same legal status as traditional paper-based processes. [20] An identity federation environment has been created to implement such a framework. The SAML v2.0 protocol has been chosen as the base communication protocol among the participants. The eIDAS SAML is an extension of the protocol, specifying communication in the cross-border eGovernment service.
The federation consists of Members States (MS) network. Its members are the states covered by the eIDAS regulation. Each of these countries

has to deploy an eIDAS Node, which acts as an identity provider for the national electronic identification scheme (eID) from any other country. All service providers must subscribe to the IdP of this country. [12] With such an approach, each citizen can be recognized across the whole network.
The protocol's extension specifies a particular format on the metadata, identifiers, attributes, format of the exchanged messages, and the process flow. According to the specification, metadata has to be structured following the eIDAS SAML Message Format document [21]. The format of the exchanged messages is defined in the eIDAS Interoperability Architecture [22].
Let us take a look at the specification of the attributes for each subject. A mandatory set consisting of a person's first name, last name, date of birth, and a unique identifier must be supported. The optional attributes available are defined based on the national law. It can include information like gender, address, birth name, or similar information. For Legal Persons, a specific set of attributes is available as well. An important thing to mention is the extensions support for providing information about identity proofing via the eIDAS Level of Assurance attribute. The complete attribute specification is referenced in the eIDAS SAML Attribute Profile. [23]

3.3.3 OpenID Connect and OAuth 2.0

Other protocols we want to describe are the OIDC[6] and OAuth 2.0. The latter of the two is an authorization framework enabling a third-party application to obtain limited access to an HTTP service, either on behalf of a resource owner by orchestrating an approval interaction between the resource owner and the HTTP service or by allowing the third-party application to obtain access on its own behalf. [24]
OIDC is a protocol built as an extensions of the OAuth 2.0. It enables clients to verify the identity of the end-user based on the authentication performed by an authorization server and obtain basic profile information about the end-user. [25]
As in SAML protocol, there are several roles taking part in the processes. The client (or RP[7]) term refers to the service, which wants to

6. OpenID Connect
7. Relying Party

access the subjects' data. The Authorization Server (AS) is responsible for authorizing the service to access the user data. Another critical role is the Resource Server (RS). It is an entity that holds the data and can provide it to the service if allowed to do so. A single entity can act both as the authorization server and the resource server. In the SAML environment, participating subjects need to store a set of descriptive data about the others. However, the metadata management process in the OIDC protocol is much simpler. The authorization server has got publicly available endpoint publishing its metadata. By processing this information, the client collects essential information about the capabilities of the server. The client application should dynamically obtain data from this endpoint rather than storing it in a hardcoded manner. To establish a connection between the client and the AS, the client must first register itself within the Authorization Server. During the registration process, the client is granted credentials for communication and specifies the communication details, like attributes that can be requested about the user or detailed information about the flow.
An essential concept that plays its part in the processes is the usage of so-called tokens. They are a set of information about the process. There are several types of them in the protocol. The most critical one is an access token. It is obtained after successful user authentication and authorization. A service can request user information from the resource server by providing the token in the request. The access token is usually valid only for a short period to increase security. It can be refreshed (or rather exchanged for a new one) using a refresh token. The second essential type is a so-called ID token. It is a piece of information provided to the end-service containing the data about the user. It provides instant access to the user information without making additional requests to the resource server.
The OIDC protocol includes a predefined set of data that the client application can request. In the protocol jargon, these categories of user data are referred to as scopes. They are further divided into so-called claims, which contain detailed information. For example, one of the predefined scopes is *profile*. It consists of several claims, like *name*, *family_name*, *given_name*, or *picture*. The standard set of scopes and claims can be extended with custom entries. The academic eduPerson [17], SCHAC [18], and voPerson [19] schemas have been adopted as well.

The critical part of the protocol is to provide the user with control over the released data. The OIDC and OAuth protocols provide a way for the user to specify what data will be released to the service. In comparison with the SAML protocol, it provides finer granularity by letting the user specify a subset of the released attributes. This granularity is usually defined at the scope level but can be further extended to the claim level. As the protocols make use of tokens to request updated user information, the important thing is to let the user revoke further possibilities to fetch the information in the future by revoking the tokens.
The main reason for developing these protocols was the need to adapt the existing processes to new trends. From the start, they have been designed and developed by big corporate organizations to suit their needs. We want to point out that they are still under development and being extended with new features. Some of them were inspirational and similar mechanisms were adopted in the SAML protocol as well.

3.4 Sample infrastructure evaluation

As we can see from the sections 3.1 and 3.2, the process of establishing a trusted identity in the target domain is quite complex. It is a crucial thing that needs to be well designed to provide a convenient experience for users and relying services. In this section, we provide an analysis of a sample infrastructure system operated in the domain of our interest. We take a look at establishing the digital identity, its representation in the environment, presentation to the services and users, provided tools for managing the identity, and documentation of the processes connected with it.
We have chosen the ELIXIR research infrastructure, and particularly the AAI it operates as an exemplary instance. This infrastructure is one of many AAIs operated in a similar environment. Some others could be the BBMRI-ERIC AAI[8], ARIA (Instruct-ERIC)[9], LifeScience AAI[10], or AAIs in a completely different environment like e-INFRA CZ[11] or EGI

8. https://bbmri-eric.eu
9. https://aria.structuralbiology.eu/
10. https://lifescience-ri.eu/
11. https://www.cesnet.cz/cesnet/e-infra-cz/?lang=en

Check-in[12]. A common thing for the mentioned infrastructures is their compliance with the AARC Blueprint Architecture [26], a document providing guidelines on designing and implementing high-quality authentication and authorization infrastructures.

3.4.1 ELIXIR AAI overview

ELIXIR is an intergovernmental organization that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage, and supercomputers. [27] The whole organization divides into several units called platforms. One of ELIXIR Compute platform goals is providing an authentication and authorization infrastructure offering a centralized user identity and access management service. The service is run as ELIXIR AAI and has been serving the organization since late 2016. The community established around it has been quickly growing since the AAI began operating. At the moment of writing, the AAI integrated more than two hundred services offered to more than six thousand users.

From the technical perspective, the AAI enables researchers to authenticate within the integrated relying services using a single AAI account. To establish the internal account, it relies on third-party identities provided by academic institutions integrated into the eduGAIN inter-federation and common social identities available on the web. Research organizations (identity providers) connect to the AAI via the SAML 2.0 protocol. Apart from the identity federation approach, digital certificates based identities can be used as well. Relying services can connect to the AAI via the OIDC or SAML interface.

The AAI provides a centralized point for user management, authentication, and authorization tools from the service perspective. As a result, the relying parties can focus more on the provided functionality rather than solving typical user-management activities. The AAI creates a unified interface offered to all of the relying parties. For example, the services can consume a predefined set of identity attributes obtained during the authentication process.

If we look at the ELIXIR AAI as a user, it creates a unified experience

12. https://www.egi.eu/services/check-in/

of using the connected relying services. It is responsible for identifying the user representation based on the external identity and performs several authoritative decisions. It provides the user with a convenient way of using their already established accounts to access the ELIXIR integrated services. Using a Single sign-on mechanism, the AAI establishes a transparent way to log in just once at a central point rather than authenticating with each of the services. It also provides a centralized point for users to manage their identities in the environment.

3.4.2 Process of establishing a trusted digital identity

A starting point for the users is establishing the identity in the infrastructure. To set it up, the user has to go through a registration process. A designated component handles all actions connected to it. To start the process, the user needs to navigate to a particular web URL that triggers authentication with the foreign digital identity. Users can either find the URL on ELIXIR webpages or receive an invitation email. After visiting the web address, the user must authenticate with an account provided by an external entity (e.g., the university). It is a transparent process of extracting foreign identity data and delivering it to the AAI. Obtained attributes are stored and used for prefilling the fields of the registration form.
As the AAI supports multiple identity providers, the user first needs to select the correct one they want to use. The environment redirects the users to a component that allows users to look for their institution in the list of integrated identity providers to perform this action. If their organization is not available, a backup solution by registering with the social identity provider has been adopted.
Following successful authentication, the user lands on a web registration form. Based on the configuration and data available from the authentication performed, form fields might be prefilled with the received information. Users can review the prefilled data and have to fill in the missing fields. At the time of writing, the initial registration requires filling in a complete person name, email address and choose a username. The last step is to read the Acceptable Use Policy (AUP) and mark its acceptance. The AUP is a document describing how the users can use the AAI as a service.
After submitting the form, the input data is stored in a separated stor-

age unit until the user verifies the email address. The process consists of visiting a uniquely generated web URL sent to the specified email address. Verification is a mandatory step to complete the registration. Behind the scenes, the registration component transforms the temporary data to an internal object describing the user and representation of the foreign identity. The identifier of the external identity is then linked to the internal user representation to create the association. The next time the user authenticates, the identifier from the identity received during the process serves for identification purposes.
The first thing we need to look at is the component for selecting the external identity provider. The interface is shown in figure A.3. It is well designed and intuitive to use. An interesting feature is the component's ability to remember the last used identity provider as shown in figure A.4. To find the eduGAIN entry for the user's organization, they have to type its name into the input box, and the component will present the matching results. Common web social identities are available via specific buttons. Integrating with backup options by using social accounts that almost every user has extends the set of users able to use the environment's services. As the authentication point varies depending on the selection made and is not in the ELIXIR AAI administration, we skip this part of the process and focus on the next step. After performing the authentication, the user needs to fill the registration form. Its design can be found in the figure A.1. Only a small set of data is necessary to establish the identity. Authentication with the foreign identity provides a user-friendly way to extract the external identity attributes. Also, prefilling the values is a nice feature to provide a better user experience. In case of all data supplied from the external source, the component automatically submits the form, and the user does not have to fill in any information. Another positive thing is informing the users about any changes in the status of the registration. Particularly, the user is informed about submitted registration, need for the email verification, successful approval of the application or its rejection.
From the negatives, visiting the registration component and sudden landing on the external identity selection page can confuse users. The places where the entry point is linked should mention that this situation will happen. An even better approach would be to land on a page that will inform the users about the procedure's details and start

the registration from this place instead. The institution selection page seems to have minor problems with internationalization, especially if the user tries to search using the name in a specific language (as shown in figure A.5). At the time of writing, the registration procedure consists of filling a single form and performing email validation. However, suppose the user would be required to go through the multiple steps (for example, data divided into categories while each category treated as a separate subform). In that case, the interface seems to lack the place for informing the user about the completed and remaining steps.

3.4.3 Consuming the digital identity as a relying party

Several services have connected to the ELIXIR AAI to provide their resources and available functionality to the ELIXIR users inside or outside the organization. As mentioned in the section 3.4.1, the AAI offers SAML and OIDC interfaces for their integration. The service has to choose the protocol that better suits its needs.
To integrate with the AAI, the service needs to register itself in the test environment. During the registration process, the service and AAI exchange metadata (or its locations), agree on the identity attributes the service will consume, and specify details of the authentication and authorization processes. The service is also required to present information about the organization providing the service, specify information like contacts, or set requirements on access management. The infrastructure provides a specialized tool that guides the representative of the service through the registration process. It also serves as a central component for managing the service settings. After successful registration and integration in the test environment, the service can request transfer to the production environment. After the transfer, it can be used by all users in the ELIXIR AAI, potentially restricted by the authorization rules.
The AAI provides a predefined set of the identity attributes released to the relying services. When connecting to the AAI, services have to determine what information they want to consume when users access the provided resources. The AAI provides a good set of attributes that include unique identifiers, user information like first and last name, username, email address, locale, organization, or country code.

Other attributes might contain values expressing the user's affiliation with the identity providers, information about user researcher status, identifiers at other research infrastructures, or information about the controlled access to datasets. Also, the AAI releases information about the identity proofing, authentication mechanisms used, or identity data freshness.
From the point of the service, it is critical to get information about identity trust level. For these purposes, several mechanisms have been adopted by the ELIXIR AAI. One of such is a standardized attribute *eduPersonAssurance* of the eduPerson schema [17]. It is an attribute presenting the service with information on the authentication process, attribute freshness, and standards met by the identity provider's management process. A secondary option is consuming a special AAI attribute representing the assurance level associated with each of the external identities. Another option is receiving additional information on the authentication process. Such an example can be informing the service about the number of factors the user has used to prove the claim on identity ownership.
As mentioned in section 3.1, the principle of minimizing the attributes set provided to the services is an essential thing. One of the requirements in establishing and using the identity from the user perspective is to minimize data released to the service while providing a sufficient amount of identity attributes to satisfy its needs. Therefore, the ELIXIR AAI informs the user about the attributes released to the service when accessing it and requiring the user to approve or reject this attribute released. In the case of a service integrated via OIDC, the user can specify a subset of the requested information that will be released. To keep the user experience smooth and not overwhelm users with making such decisions, they can decide that the environment should ask them to do it only if the requested attribute set changes. The interface is shown in the figure A.2.

3.4.4 Digital identity representation

An important thing is how the identity is presented both to the users and the relying services. In this example environment, the user identity consists of two datasets. The first type holds information about the identity as received from the external source. The second type

represents the internal user account. It holds information managed in the AAI and attributes derived from the external identities information. These two entities are linked together and form the complete user identity.

We have already stated that the environment relies on external academic identities or digital certificates. Such an identity provider needs to release several attributes to integrate into the ELIXIR AAI. Namely, it has to forward a unique identifier of the identity (*eduPersonUniqueId*), the principal's name (*eduPersonPrincipalName*), the user's affiliation with the identity provider (*eduPersonScopedAffiliation*), and the identity provider's domain name (*schacHomeOrganization*). Additional attributes can be released as well, but the mentioned set is the minimum needed. The latter three attributes can be accumulated from multiple user identities and released to the services. The identifiers of identity and identity provider are used to recognize the internal user representation after successful authentication. Other attributes can be stored and later used as partial information to form the identity in the environment. For the identifier, the AAI creates its unique identifier for each user, and it is the primary identifier released to the services.

An important thing for the AAI is to learn the trust level for a particular user identity. As mentioned in the section 3.4.3, ELIXIR AAI has integrated several frameworks for indicating such information to the relying services. However, it also needs to learn this information from an external source. The first way of doing so is by consuming the SAML eduPersonAssurance attribute from the identity provider. The secondary approach is storing a so-called Level of Assurance for each of the identities, based on the forwarded value or verification processes performed with the user's cooperation (i.e., validating email address or performing multi-factor authentication).

A central point for the users to review the identity attributes is an application that serves as a user profile. It lets the users review the data the AAI stores about them in original identity and derived attributes. It also lets the user set preferences in the environment, like preferred localization for the interfaces or email addresses. Apart from that, the profile page provides links to additional services that can be used for displaying and reviewing attributes that might get released to the end service but are generated dynamically at the time of accessing the service. In the recent changes, the profile page added support for

setting up multi-factor authentication methods.
The user interface shown in the figure A.6 depicts the mentioned user profile. The application needs the user to authenticate using the AAI account. It is designed in a modern way, using appropriate tools. After the users log in, they are presented with an initial overview of the identity. The attribute set is consolidated to the most important ones, like the user's name, affiliations with identity providers, or login and email address. Via the menu, users can switch to view the detailed information about the particular external identities. Under the Privacy tab, the user can review the information that each of the structural units and the whole environment holds about the subjects. Via the settings subpage, the users can manage their preferences like the mentioned multi-factor authentication methods.

3.4.5 Tools for identity management

When the identity already exists, it is essential to have tools for its management. Such activities include updating the attributes to up-to-date values, removing the obsolete information, adding new data, or modifying the information based on the user preferences. ELIXIR AAI provides several tools for performing such tasks.
The first option is an automated mechanism triggered at authentication to the AAI with the external identity. When the user authenticates, and the AAI receives the data from the identity provider, it updates the internal representation with the new values. The derived attributes based on the stored identity data are automatically recomputed and updated as well. This mechanism provides a transparent way of updating the identity attributes. However, there are some flaws connected to this approach. For example, when the user gains a new affiliation with the identity provider, the AAI is not informed about it until the authentication happens. Due to the Single Sign-On mechanism, the event can be even further delayed. As users can link multiple identities to a single AAI account, they often link an identity established at the social identity provider. Many of them often have already authenticated with such an account and prefer to use it to access the AAI relying services. Thus, the AAI might not receive information about the new affiliation at a sufficient time window. The services need to be informed about the freshness of released data to prevent such situations. This problem

can be easily solved by releasing additional attributes.
The second way is using the identity management system. This approach is intended to be used by the AAI operators. In the ELIXIR AAI, a tool named Perun [28] developed by CESNET, and ICS MUNI[13] is used for these purposes. Perun covers the management of the whole ecosystem around the user's identities, groups, resources, and services. In ELIXIR AAI, it is responsible mainly for three tasks. The first of them is the management of the user identities and user accounts. The second task is providing abilities for a user classification within the structural units (groups and virtual organizations), which is a vital task for extending the identity data. For example, information about membership in different groups can be used for authorization decisions and released to the relying services in specific attributes (e.g., *eduPersonEnitlement*). The third activity is providing the user-related data to the services via a provisioning mechanism.
The last option we want to mention is designed mainly for the users. The AAI provides a specific tool serving as a user profile, which has been described in section 3.4.4. As already mentioned, the application allows users to review the stored identity data, manage preferences, and link or unlink other external identities to the AAI account.

3.4.6 Documentation of identity-related topics

An essential part of the environment is documentation. It serves as an information source for the users or consumers of the provided functionality (e.g., the relying services). It is crucial to provide these subjects with such places to explain the infrastructure processes, inform them about the available functionality, explain the interfaces, structure of the AAI, and similar things. From the user's perspective, they need to be informed on the process of establishing the AAI account, the way of using it and benefits it provides, consumed external identity attributes, data usage, and data protection and processing policies. If we take a look from the perspective of a relying service connected to the AAI, it needs to know how the authentication and authorization processes work, functionality that can be leveraged to the AAI, available user attributes the service can consume, or processes

13. Institute of Computer Science, Masaryk University, Brno

of integration into the environment.
From the user documentation, the first important thing is informing the users about the possibility of establishing the AAI account and the reasons why users might be interested in doing so. Such information is available on the ELIXIR webpages under the AAI section. It provides extensive information on the registration process. It also describes the functionality the account grants the user and the process of using it. However, this information may appear hidden in the organization's web pages. FAQ[14] section is an excellent way of communicating the most common topics users may need to discuss. These pages also link to the data usage and privacy policy documents, a description of the AAI from a technical perspective, and contact information. The secondary page that might be a solid place to look for information is the AAI homepage. It contains pointers to different vital services and pages. Unfortunately, this page seems also to be lacking references from external sources and thus might be hard to find.
If we now switch to the role of a relying service, probably the most important thing is the documentation on how we can integrate to the ELIXIR AAI. This process is well described in a document linked from the AAI's official web pages. It points to many other documents describing the infrastructure. One of such is the description of user-related attributes the service can consume. The documents are well structured and highly cross-referenced. However, the document-like approach might not be the most user-friendly way, and a wiki collecting all the information in one place could be a better solution. The identity's life cycle does not seem to be documented too much. We want to point out a brilliant idea of setting up a video platform account where screencasts, webinar recordings, tutorials, and manuals can be published. The AAI is also making use of the ELIXIR e-learning platform, where it publishes training materials. This service is commonly known to ELIXIR users.

14. Frequently Asked Questions

4 Guidelines

Building an infrastructure to suit the requirements set in section 1.4 is not easy. This process involves many decisions with many options from which to choose. This section provides several tips to help readers sort out the options and make the design process more manageable.

4.1 Designing the environment

In this part, we focus on the decisions in designing the environment as a whole. We first discuss the topic of choosing the appropriate types of external digital identities the environment will use. From the identity attributes, one of the most important is identity trust. We debate mainly the framework for representing this property and expressing it to the relying services. The next topic we consider as critical is interconnecting the different components.

- Select external digital identity types that the infrastructure will integrate with.
- Pick a mechanism for representing the identity trust.
- Design the general architecture. Pick components that will form it. Consult already existing guidelines for designing and interconnecting the components.

4.1.1 Choosing the correct external digital identity types

In section 2 we have analyzed different types of digital identity. From the evaluation performed, we have selected the federated identity (section 2.3.1), digital certificate-based (section 2.7), and the electronic identification (section 2.9) as the most suitable options for the purposes of the selected domain. However, other mentioned types might be applicable as well depending on the more specific requirements of the particular situation.
The federated identity has been evaluated as the best choice. We recommend this approach if the users come from a common domain

that might form a federation of identity providers. It is also applicable on an international scale. Good examples are the academic federations described in section 2.3.1. Implementation is potentially low in effort and brings high value to the environment. Environment's entry point has to act both as service and identity provider via a component referred to as proxy. This component serves as a hub point for integrating relying services and identity providers. Information value is at a sufficient level, and the plus side is the predefined attribute schemas that can be used. This approach has quite good support for expressing the identity trust levels. A disadvantage might be the need for a specialized component that allows the user to select the particular identity provider from the integrated options. This additional step in the authentication process might leave users confused and needs to be properly explained. Another tedious task can the federation (both identity provider and relying parties) metadata management. Overall, this approach can be quickly built and put into production.
The electronic identification described in section 2.9 has been marked as the second most suitable option in our evaluation. We recommend picking this choice when users' identity proofing is critical. Integrating this identity type should not be difficult. The approach of using a proxy (serving both as identity and relying service) might be applicable in this scenario as well. A tricky part might be on the formal side as the environment needs to connect to a local eID identity provider. Another difficulty might be the international nature of the LS community infrastructures. The cross-border identification with eID is, at the time of writing, not at the desired level. However, the standard attributes set, the low possibility of forging the identity, and exceptional trust level outweigh the mentioned difficulties. Also, we suppose a significant development will happen in this area. One more problem might be caused by the need for specialized hardware on the side of users. However, this situation will be solved with the eID becoming more popular and used by the users for their primary purposes (eGovernment and similar tasks). Also, the hardware keeps being developed, and we suppose that specialized devices will not be necessary for future use.
The last selected option has been the identity based on the digital certificate. This solution might be optimal if users come from different domains that cannot be easily joined together, i.e., forming an identity federation. The vital thing to decide is the set of certificate authorities

that should be considered as trusted. A strong side of this type is the availability of the users, as anyone can get the certificate. We also suppose this approach will gain more popularity due to tasks like document signing moving towards a digital alternative. The flaw of this approach that needs to be taken into account is the possibility of the credentials being compromised and lower data freshness compared to the previous two solutions. On the other hand, it might offer lower interaction needed from the user perspective and result in a better user experience. For example, the users could import the certificate into their web browser and be required to select the correct certificate for authentication instead of using credentials set or authentication tokens.
An important thing to keep in mind is providing a backup solution for the users that might not have access to the chosen identity sources. In some cases, they might not be allowed to use their identity, e.g., due to law regulations or their privacy concern. In such a situation, the environment should provide an alternative solution for these users. A good choice might be integrating with social network identity providers or creating a possibility to set up and use a local account. The important thing is to represent lower identity trust in such cases or performing the verification process.

4.1.2 Identity trust representation

One of the essential concepts connected to digital identity is the identity trust level described in section 1.3. From the perspective of designing the environment, we consider the following three tasks as critical. The first is deciding the minimal level of trust the identity has to achieve to be considered usable. In a scenario, as we have described in section 1.4, we recommend using a formally verified identity (e.g., by verification using an identification document during the establishment process). All of the types of existing digital identities we have selected as applicable in section 4.1.1 can satisfy this requirement. An additional verification might be performed at the time of establishing the association with the external identity. It can be done, for instance, by requiring the user to visit a uniquely crafted web link delivered via email, sending a one-time password to a mobile phone, or similar mechanisms. We strongly advise implementing at least one of such

methods. Another solution might be the requirement of using multiple tokens during the authentication. A good source of information can be the AARC's Recommendations on Minimal Assurance Level Relevant for Low-risk Research Use Cases [29].
The second task is the choice of how the identity trust will be represented. We strongly recommend using one of the existing frameworks. For the federated identity, a good choice is the already mentioned REFEDS assurance framework [6] or creating a designated attribute containing the level of assurance. With the electronic identification, we suggest using an approach similar to the one developed in the eIDAS SAML3.3.2. An identity based on digital certificates might extract this information from the certificate itself, as it may contain information on the assurance level. A necessary task is to distinguish between identity trust levels among different identity providers and identity types in the case of integrating multiple options. A mapping between the integrated solutions should be established and published as a resource for the relying services (e.g., the highest assurance in the federated identity might not be equal to the highest value in electronic identification). A good source of information for implementing such a framework or its selection might be the ISO/IEC 29115:2013 [7] document. We also recommend studying the Comparison Guide to Identity Assurance Mappings for Infrastructures (AARC-I050) [30].
The third activity is deciding how the identity trust level will be represented to the relying services. We strongly recommend using a mechanism available in the selected protocols for communication with the relying services if possible. For instance, in the SAML protocol, we recommend using the authenticationClassReference property for expressing the method of authentication used. We suggest using the REFEDS assurance framework [6] or a similar mechanism for providing data about identity verification. The mentioned solutions are also available in the OpenID Connect [25] protocol. AARC project has published a set of guidelines on exchanging the identity trust information in the document Guideline on the exchange of specific assurance information between Infrastructures (AARC-G021) [31]
An important topic to consider is the scenario when users are allowed to combine their identities. In such a case, the attributes are combined from different identity sources. Furthermore, the identity trust level has to be reevaluated based on the merging strategy. We advise finding

more details in the AARC's Account linking and LoA elevation use cases, and common practices for international research collaboration [32].

4.1.3 Designing and building the environment's components

So far, we have chosen what digital identities we will rely on and how the identity trust will be represented. When designing the environment components, we strongly recommend studying the AARC Blueprint Architecture [26]. The Blueprint Architecture lets software architects and technical decision makers mix and match tried and tested components to build customized solutions for their requirements. [26] The document is intended to be used for a federated environment, but the concepts might be elsewhere as well. As visible from the figure A.7, it already considers more than just federated identities and even combining them.
Probably the most important of the components that need to be included is the place for relying services integration. In the AARC BPA [26], this component is referred to as Proxy. Together with the Discovery Service, which allows the user to select the identity provider from the list of integrated options, these are the first components the users interact with. We recommend putting significant effort into designing and implementing them.
The next pieces of the environment are the relying services. From the building of the environment tasks, the providers need to implement the interfaces to allow services to connect and consume user attributes. We advise using the protocols mentioned in section 3.3. If possible, we strongly recommend using both of them. The main reason is the wide variety of services that will be able to connect to the environment.
As well as the AARC BPA [26] advises, we recommend implementing components that will handle things like displaying AUPs to the users and requiring them to accept such documents and terms. Also, the environment should be able to integrate with data repositories originating from outside of the identity providers. These services can either connect via the interfaces provided for relying services or be directly integrated into the environment via specific interfaces serving such purposes.

4.2 Creating the association of external identity with internal user representation

- Define required information that needs to be obtained during the process.
- Design the process of establishing the identity. Decide the flow and try to minimize the steps needed.
- Make decisions on the process of prefilling the inputs and validation.
- Decide mechanism for providing notification for the user.

The first process users experience when reaching the environment is the identity establishment. As described in section 3.1 and sample evaluation, this process needs to be well designed, guide the user through it, consist of the smallest number of steps possible, and require as few inputs from the user as possible. The steps should be transparent and automated. If user interaction is needed, the involved person should be provided with information on what is happening and what tasks need to be done.
We distinguish two main flows for the user to reach the point of establishing the community infrastructure's identity. The first one is going through a designated flow. This flow ends when the identity is established. The second way is a result of an activity that triggers the identity creation as a side outcome. An example could be the user visiting the service and performing authentication. Instead of ending in a scenario when the service states the user has not been recognized, the user will end authenticated at the service with the identity establishment happening in the background. It can also result in the user going through the primary registration flow with additional returning to the original service. We strongly recommend implementing both options of establishing the identity.
Let us now focus more on the establishment process itself. As we have already stated, it should consist of the minimal number of steps possible. The goal of the process is to receive necessary information from the user and deliver it to the target environment that can use it (e.g., the relying service accessed by the user). An example could be filling a

simple web-based registration form. Because of relying on third-party existing user identities, the critical part is extracting the external identity's identifier. The process should be automated as much as possible. Our advice is to require authentication with the existing foreign identity. This activity can serve as a method for automated extracting data from the external identity source. Obtained data can then be used for prefilling the web-form fields. If all the input fields can be satisfied with the extracted data, the form could be submitted without further user interaction. A considerable step is letting the user validate or change the prefilled values. This approach's positive is the possibility of incorrect or outdated information being updated at the stage of identity establishment. As for downsides, we consider the need to validate further data updated by the user. Another disadvantage is the need to decide the information that the user can modify (e.g., username can be modified, but the external identifier should not). Last, and probably the most serious, is the requirement of additional interaction from the user side, making the process less automated. If the process requires the user to going through several steps, like filling multiple forms, we strongly recommend consolidating these forms into a single instance. The important thing is trying to minimize the amount of data the user is required to provide while being sufficient to obtain the required information for the needs of the environment.
A critical step is the validation of the provided data. In case of requiring information that needs to be unique, we advise implementing an automated process triggered by each change of the particular value to check for the availability and conflicts with existing values. We recommend implementing a mechanism for validating information like email addresses, mobile phone numbers, and similar. Such verification can also be considered as a multiple-factor authentication method and therefore increase the trust in the provided data.
Other important parts of the process are notifications sent or displayed to the user. They can provide explanatory texts about what processes are happening in the background, provide error information and guides on how to resolve them, or inform about the establishment process's status. Notifications can be displayed directly in the registration interface, especially if they inform about the particular events that have occurred and it makes sense to respond to them. Otherwise, we advise providing information in an asynchronous manner, e.g.,

via email messages about any events that occur, and the user should be informed about them. An example can be an email notification about the approved registration if it needs manual processing by a designated person after being submitted.
In an optimal solution, the user must not fill any data or visit any specific component to trigger the registration process. The user starts by accessing the service they want to use. Then, the service triggers the authentication process. The user continues with performing the authentication with the third-party identity. In the background, the identity in the target environment is established. During the process, the user is not required to interact in any other way. The process finishes with the user being authenticated with the service. Instead of requiring the user to fill in missing data or correct it, the user can be later required to visit a designated component where these tasks can be performed.

4.3 Presentation of a digital identity to the user and relying party

- Provide a specific component to the users for presentation of the identity.
- Choose a correct representation of identity attributes.
- Define a representation of the identity attributes for relying services.

In this section, we assume the identity has been successfully established and can be used, e.g., to access the relying services. The identity itself might be transparent for the users. However, we strongly advise including a component responsible for presenting the internal user representation to its owner. In particular, such a component should inform the owner about the data associated with them and include a representation of the external identity. An example of such a component might be an application serving as a user profile.
When designing it, a critical task is deciding how the data will be presented to the identity owner. Several attributes might contain values with no straightforward meaning for the user. For example, the

attribute in which the user name is stored can be displayed without modification because of the value's straightforward semantics. However, attributes representing information like the level of assurance, or identifiers, might not be easily understood by the user. Some of the attributes do not have to be displayed at all, as it has no information value for the user and might have a disturbing or confusing effect. As a result, we advise designing the component so that the user is first presented with the most important information, like a primary identifier, name, email, gender, localization, and similar. The attributes should be presented with an understandable name instead of name of the attribute from the used frameworks (e.g., *givenName* identified as *urn:oid:2.5.4.42* in SAML v2.0). The presentation should also include a description of what the attributes represent, in some cases including an example value. For attributes with difficult value semantic, a mapping should be used for presenting the value. An example can be expressing the level of assurance defined in the ISO/IEC 29115 Standard [7]. Instead of using the values set by the standard - LoA1, LoA2, LoA3, LoA4, a description of the value should be used instead. In some cases, using visual representation, e.g., icon or image, can be used alternatively, especially if it can replace long textual description. Another important thing is deciding how the identity will be represented to the relying services. It heavily depends on the framework and protocols used to communicate with the services (e.g., the SAML v2.0 or OpenID Connect described in section 3.3). Many already include predefined ways of representing certain information or can be further extended. Especially in the latter case, documentation on the particular attribute syntax and semantics must be established and made available to the service developers. The infrastructure should consult the needs of the services and provide data to satisfy them. We strongly recommend adopting already existing protocols for communication and predefined frameworks for information representation.

4.4 Life cycle of a digital identity

- Design a lifecycle of the identity - its establishment, use, management, and the end of life.

- The lifecycle should be documented and in some form provided to the users and relying services.

From the infrastructure point of view, a crucial task is defining the lifecycle of the identity. It has to cover different stages of using the identity. The first step is its establishment. Following that, the identity enters the stage of being used and managed. After some time, when the digital identity becomes obsolete for different reasons, it should be decommissioned and reach the end of life stage.
The stage of identity establishment corresponds to the initial authentication or registration process. In this stage, the community infrastructure should create the internal representation of the identity. For these purposes, it needs to use the attributes of the external digital identity the environment relies on, and the user wants to use. A more detailed description of the process has been provided in section 4.2.
The next stage is using and managing the identity. During this phase, the user can use it to authenticate to relying services. An important task is the management of identity. It includes updating the attributes to up-to-date values, removing obsolete information, and enriching the information set with new data. Such actions can be performed by the infrastructure operators, automatically b the underlying systems, or manually by the user. In the first scenario, the automatic procedure should be invisible to the user. A simple mechanism we advise to use is updating the data whenever the user authenticates, as the authentication provides an infrastructure with a fresh set of data. From the operator's perspective, this can include making manual changes in the user's attributes or managing the user membership in structural units. For example, adding the user to a specific group should be reflected in attributes based on this kind of relationship. The second way of updating the data is manually by the identity owner. We suggest creating a component for performing such tasks. It can and should be consolidated with the component for presenting the user's identity described in section 4.3.
While using the identity, the owner has to be informed about the associated data. Apart from designing a component for presenting the identity to the user (as described in 4.3), they should also be made aware of what data is being released to the relying services. For these purposes, the protocols described in section 3.3 include the concept of

giving consent to forwarding the attributes to another party. Using such functionality is mandatory from our point of view. However, we advise including a possibility for the user to opt-out of being presented with the request to approve the data transfer each time they authenticate to the relying service. A *remember* function should be available for such purposes. In case of changes in the attribute set requested, the user should be prompted regardless of the previous opting-out. We suggest prompting the user for consent at least once during a predefined time window, even if the attribute set does not change. It will help users understand that their data is still released to a third party, and they can take additional actions if they do not want to allow it. Also, the possibility to remove the opt-out decision should be available. For internal components that require the identity data, the consent does not need to be enabled as the data remains used in the same environment.

The final stage of the lifecycle should be the identity decommission. It can happen as a result of the user's decision to stop using the infrastructure. Also, the identity should be made obsolete when the user has stopped using it for a potentially long time (e.g., one year). In case of such an event, the user should be informed about the fact and granted a grace period, during which they can prove the data is still needed and the account will be further used. After the grace period expires, the decommissioning process should start. From the internal point of view, the user representation should be deleted or anonymized as much as possible. Critical information that should not be further reused by any other user, like unique identifiers, can be kept for such purposes. However, it needs to be disassociated from the particular user and become non-identifying. From the external perspective, the relying services need to be informed about the identity decommission and take actions corresponding to this event. The infrastructure should enforce the service's behavior in such a case by an agreement between the participating parties. The whole stage and processes it implies need to be documented for internal (environment operators, helpdesk) and external (users, relying services) subjects.

4.5 Documentation of identity-related topics

- Provide user documentation - identity description and establishment, using the identity, data processing, and legal aspects.
- Create relying service documentation - communication protocols supported, identity attributes available, integration process.
- Establish the Identity lifecycle documentation, Privacy policy, Acceptable use policy.

The last topic we want to discuss is the documentation provided to the users and relying services. The documents need to be publicly available and capture detailed information while keeping the text as straightforward as possible.
From the user's point of view, the first document that needs to be available is the identity's description in the infrastructure. The document needs to contain information about the identity in the environment, how it can be used, and what benefits it brings to the user. It should also include a description of the detailed establishment process and using the identity in the environment. Other essential documents for users are the legal aspects and information on data processing. Such documents should include a description of how the identity attributes are further processed, released to the services, enumerating and explaining what and why data needs to be collected.
If we consider the service's perspective, the infrastructure needs to provide primarily technical documents on the integration and its legal aspects. In the technical documentation, the service should learn how the infrastructure is built, what communication protocols and interfaces it supports, and the integration process description. It should include the steps the service developer needs to take to integrate the service into the infrastructure successfully. A part of this must be the documentation on what identity attributes the service can consume, syntax, and semantics of the attribute values.
Some other documents might be interesting for multiple subjects. For example, documentation on the identity lifecycle should be available for both users and relying services. It can be written as multiple separate documents, each focusing more on the particular perspective. We

strongly advise creating such a document for these purposes. Besides, while creating it, the infrastructure designers might often find out caveats that might occur during the transition between different life cycle phases.

5 Conclusion

Digital identity is an essential concept in the virtual environment. The particular domain in which the identity is used has a significant influence on the required properties of the identity. It influences what data it contains, in what amount, how well it has to be verified, or how it is used. Many people have already established several digital identities. Instead of creating new ones, already existing identities should be reused and modified according to the particular concept.
In section 1, we have introduced the concept of digital identity. We have provided its definition, discussed security aspects, and described the processes connected to it. An important attribute of a user identity in the digital world is the identity trust, which we have discussed in section 1.3. We have also discussed a digital identity in the domain of life science research infrastructures.
Due to the international character of such infrastructures, they should rely on already existing digital identities instead of creating new ones on their own. In section 2, we describe several types of digital identities which can be integrated into such environment and used for user identification. We have briefly described each of the types, considered its strong and weak sides, and evaluated it according to the before-set needs of the chosen domain. At the end of this section, we have cross-evaluated all of the mentioned types and chosen the most appropriate options that we consider as applicable for such an environment.
In section 3, we have focused on a process of creating a person's representation using one of the selected types of digital identities in the environment of life science infrastructures. We have described the process both from the user and relying service perspective. After that, we have discussed suitable communication protocols that can be used for integrating the relying services and connecting with external identity providers. In the last sub-section of this part, we have analyzed the ELIXIR AAI infrastructure. Particularly, we have provided an introduction into the ELIXIR AAI infrastructure, analyzed the process of associating the user representation, its consumption from the perspective of relying service, available tools for managing the identity, and the available documentation on such tasks.
The final section 4 provides guidelines for establishing an environment

similar to the life science research infrastructures. We have provided advice on what components to use for building the environment, how to choose the appropriate type of external digital identities to rely on, and how to represent the concept of identity trust. We have also given tips on the process of associating the external identity with the internal user representation. The other guidelines are aimed at the area of representing the digital identity to its owner and relying on services, its life cycle, and the needed documentation.

Bibliography

1. LOPEZ, J.; OPPLIGER, R.; PERNUL, G. Why have public key infrastructures failed so far? *Internet Research*. 2005, vol. 15, pp. 544–556. ISSN 1066-2243. Available from DOI: 10.1108/10662240510629475.
2. WINDLEY, P. J. *Digital identity*. 1st. Farnham: O'Reilly, 2005. ISBN 9780596008789.
3. ANDRESS, J. *The Basics of Information Security*. The basics of information security: understanding the fundamentals of InfoSec in theory and practice. 2nd. Syngress, 2014. ISBN 9780128007440.
4. GRASSI, P.; GARCIA, M.; FENTON, J. *Digital Identity Guidelines*. 2017. Available from DOI: 10.6028/NIST.SP.800-63-3.
5. THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. *Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC*. 2014. Available also from: https://eur-lex.europa.eu/eli/reg/2014/910/oj. Accessed on 2021-03-21.
6. REFEDS. *REFEDS Assurance Framework*. 2020. Available also from: https://refeds.org/assurance. Accessed on 2020-12-9.
7. *Information technology – Security techniques – Entity authentication assurance framework*. 2013-03. Standard. International Organization for Standardization.
8. THE EUROPEAN PARLIAMENT, AND THE EUROPEAN COUNCIL. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. Available also from: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

9. THE PARLIAMENT OF THE CZECH REPUBLIC. *The Act No 181/2014 Coll.* 2014. Available also from: https://www.nbu.cz/cs/pravni-predpisy/1091-zakon-o-kyberneticke-bezpecnosti-a-o-zmene-souvisejicich-zakonu-zakon-o-kyberneticke-bezpecnosti/.

10. THE NATIONAL COUNCIL OF THE SLOVAK REPUBLIC. *The Act No 69/2018 Coll.* 2014. Available also from: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2018/69/.

11. HADWICK, D. W. Federated Identity Management. In: *Foundations of Security Analysis and Design V: FOSAD 2007/2008/2009 Tutorial Lectures*. Ed. by ALDINI, A.; BARTHE, G.; GORRIERI, R. Springer Berlin Heidelberg, 2009, pp. 96–120. ISBN 9783642038297. Available from DOI: 10.1007/978-3-642-03829-7_3.

12. CARRETERO, J.; IZQUIERDO-MORENO, G.; VASILE-CABEZAS, M.; GARCIA-BLAS, J. Federated Identity Architecture of the European eID System. *IEEE Access*. 2018, pp. 1–1. ISSN 2169-3536. Available from DOI: 10.1109/ACCESS.2018.2882870.

13. ULRICH, A.; HOLZ, R.; HAUCK, P.; CARLE, G. Investigating the OpenPGP Web of Trust. In: ATLURI, V.; DIAZ, C. (eds.). *Computer Security – ESORICS 2011*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 489–507. ISBN 9783642238222.

14. WINNARD, K.; BUSSCHE, M.; CHOI, W.; ROSSI, D.; REDBOOKS, IBM. *Managing Digital Certificates across the Enterprise*. IBM Redbooks, 2018. ISBN 9780738441504.

15. ESTONIAN INFORMATION SYSTEM AUTHORITY. *ID.ee*. 2021. Available also from: https://www.id.ee/en/. Accessed on 2021-3-23.

16. OASIS. *OASIS Open*. OASIS, 2020. Available also from: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security#overview. Accessed on 2020-12-28.

17. REFEDS. *eduPerson and eduOrg Object Classes*. 2021. Available also from: https://refeds.org/eduperson. Accessed on 2021-2-21.

18. REFEDS. *SCHAC - SCHema for ACademia*. 2021. Available also from: https://wiki.refeds.org/display/STAN/SCHAC. Accessed on 2021-01-05.

19. VOPERSON. *voPerson*. 2020. Available also from: https://voperson.org/. Accessed on 2020-10-17.

20. CUIJPERS, C.; SCHROERS, J. Eidas as guideline for the development of a pan European eid framework in futureid. In: HÜHNLEIN, Detlef; ROSSNAGEL, Heiko (eds.). *Open Identity Summit 2014*. Bonn: Gesellschaft für Informatik e.V., 2014.

21. EIDAS EID TECHNICAL SUBGROUP. *eIDAS SAML Message Format version 1.2*. 2019. Available also from: https://ec.europa.eu/cefdigital/wiki/download/attachments/82773108/eIDAS%20SAML%20Message%20Format%20v.1.2%20Final.pdf?version=3&modificationDate=1571068651727&api=v2. Accessed on 2021-3-13.

22. EIDAS EID TECHNICAL SUBGROUP. *eIDAS Interoperability Architecture version 1.2*. 2019. Available also from: https://ec.europa.eu/cefdigital/wiki/download/attachments/82773108/eIDAS%20Interoperability%20Architecture%20v.1.2%20Final.pdf. Accessed on 2021-3-13.

23. EIDAS EID TECHNICAL SUBGROUP. *eIDAS SAML Attribute Profile version 1.2*. 2019. Available also from: https://ec.europa.eu/cefdigital/wiki/download/attachments/82773108/eIDAS%20SAML%20Attribute%20Profile%20v1.2%20Final.pdf?version=2&modificationDate=1571068651772&api=v2. Accessed on 2021-3-13.

24. HARDT, D. *The OAuth 2.0 Authorization Framework*. 2012-10. Tech. rep. Available also from: https://tools.ietf.org/html/rfc6749.

25. SAKIMURA, N.; BRADLEY, J.; JONES, M.; MEDEIROS, B. de; MORTIMORE, C. *OpenID Connect Core 1.0 incorporating errata set 1*. OpenID Foundation, 2014-11. Tech. rep. Available also from: https://openid.net/specs/openid-connect-core-1_0.html.

26. AARC CONSORTIUM PARTNERS AND APPINT MEMBERS. *AARC Blueprint Architecture*. Ed. by LIAMPOTIS, N. 2019. Available also from: https://aarc-project.eu/architecture/. Accessed on 2021-01-03.

27. ELIXIR. *ELIXIR Web - About us*. 2021. Available also from: https://elixir-europe.org/about-us. Accessed on 2021-3-12.

28. MASARYK UNIVERSITY BRNO AND CESNET, Z. S. P. O. *Perun*. 2021. Available also from: https://perun-aai.org/. Accessed on 2021-03-13.

29. LINDEN, M.; GROEP, D.; PÖHN, D.; COULOUARN, T.; PEMPE, W.; SHORT, H. *Recommendations on Minimal Assurance Level Relevant for Low-risk Research Use Cases*. 2015. Available also from: https://wiki.geant.org/pages/viewpage.action?pageId=123765209&preview=/123765209/123769445/MNA31-Minimum-LoA-level.pdf.

30. GROEP, David L.; NEILSON, Ian. *Comparison Guide to Identity Assurance Mappings for Infrastructures*. Zenodo, 2019. Available from DOI: 10.5281/zenodo.3627594.

31. CONSORTIUM, AARC; MEMBERS, AppInt. *Guideline on the exchange of specific assurance information between Infrastructures (AARC-G021)*. 2018. Available from DOI: 10.5281/zenodo.1173558.

32. PARTNERS, AARC Consortium; MEMBERS, AppInt. *Guidelines for the evaluation and combination of the assurance of external identities (AARC-G031)*. 2018. Available from DOI: 10.5281/zenodo.1308682.

A Figures

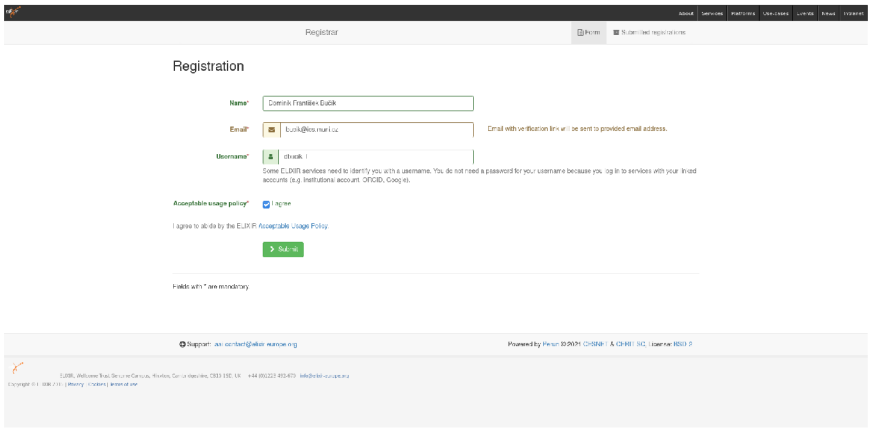

Figure A.1: Registration web form for establishing an ELIXIR AAI account

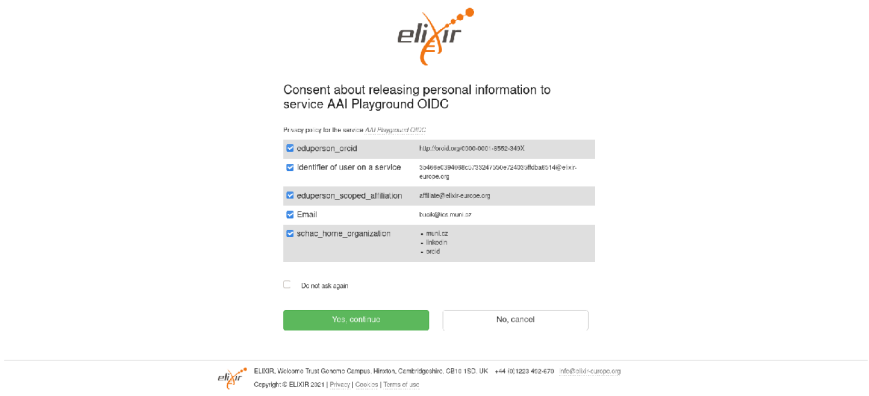

Figure A.2: Interface for approving attribute release to relying service

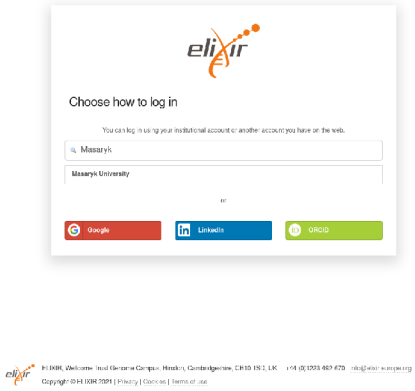

Figure A.3: External identity provider selection page

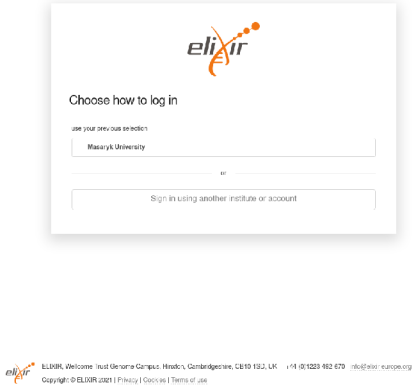

Figure A.4: External identity provider selection page - previous selection

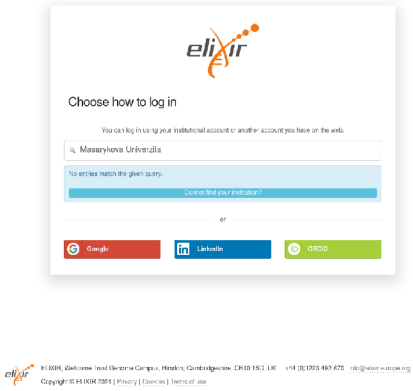

Figure A.5: Language specific issues in the identity provider selection page

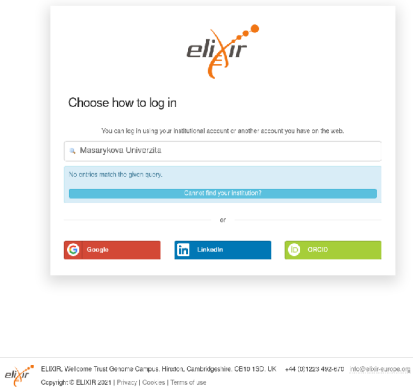

Figure A.6: ELIXIR AAI User profile application

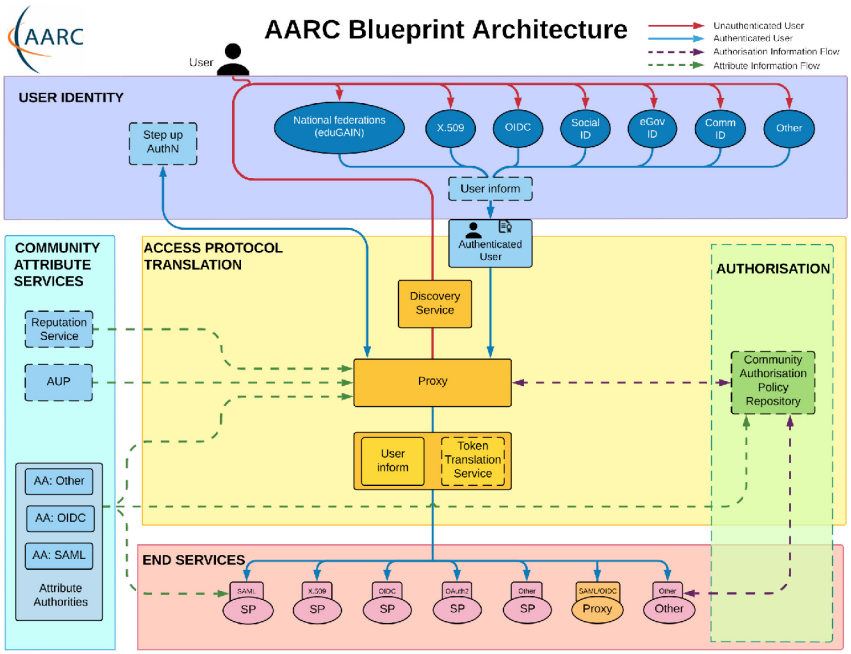

Figure A.7: AARC Blueprint Architecture [26]