15-06-2016

Milestone MJRA1.2: Design for Deploying Solutions for "Guest Identities"

Milestone MJRA1.2

Contractual Date:	30-04-2016
Actual Date:	15-06-2016
Grant Agreement No.:	653965
Work Package:	JRA1
Task Item:	JRA1.3
Lead Partner:	STFC
Document Code:	MJRA1.2
Authors:	J Jensen (STFC) (ed), U Stevanovic (FZJ), I Kakavas, N Liampotis, C Kanellopoulos (GRNET), M Haase, P Gietz (DAASI), M Jankowski (PSNC), M Reale, M-L Mantovani (GARR), L. Florio (GÉANT)

The research leading to these results has received funding from the European Community's Horizon2020 Programme under Grant Agreement No. 653965 (AARC).

Abstract

This document looks at the advantages and risks associated with guest identities, as well as the practical aspects of using them in e-infrastructures. Guest identities are defined here as identities used outside of their original context, be they social media IDs, bank or government IDs, or identities provided by research communities that were originally intended for use only within that community.

Table of Contents

Table of Figures

Table of Tables

Executive Summary

This document provides information about deploying and using "guest identities," information that in turn feeds into the blueprint architecture document (M15). The document also offers guidance on this topic to EU-wide infrastructures and collaborations.

The premise behind much of the work of AARC is that users should be able to authenticate to e-infrastructures and research infrastructure using identities they already have. While the infrastructure could safely use identities from users' "home" institutes – employer, university, etc. – and national access management federations, this document focuses on other, sometimes less formally defined, identities: social media, government IDs, IDs provided by research communities, etc., which are referred to here as "guest identities." Using guest identities, infrastructures can thus broaden their user base, particularly to the so-called "long tail" – members of the public who contribute to research – or to users who belong to research institutions that are not part of identity federations. Guest identities can also lower the barrier for new users. There are plenty of examples where using guest identities is advantageous and where the level of assurance is good enough, provided the risks can be managed. This document discusses these risks and benefits.

A key question is the level of assurance (LoA) assigned to the identity, or assigned to its use in authentication. Comprising multiple factors, including how the user's identity was checked when the identity was created, the LoA can be used by the resource to decide whether to rely on the identity for a particular purpose. Some uses, such as account presentation, typically need not have a high LoA; in contrast, billing or access to sensitive data often require a high LoA.

Guest identities are defined here as *identities used outside of their original context*. Different types of guest identities are considered in this document:

- Social media IDs – Google, Facebook, Microsoft Live, Linkedin. The advantage of using these identities is that users typically already have one or more of these, and, as highlighted by the FIM4R study, younger researchers will expect to be able to continue to use these identities to carry out their research (and blog about it). There is, however, a strong demand for researchers to also be able use these IDs to access research facilities, and social media identities may be sufficient for a wide range of work.
- Government or banking IDs – the availability and reusability of these identities is typically bank/country-dependent, but they generally provide a very high level of assurance. For international collaborations, the adoption of these identities is not very common and at the moment not very scalable: In the case of bank "eIDs," there may be national infrastructures that can use the IDs directly. However, for international projects, a resource provider would have to trust many different banks, which would require effort to manage the process. Government eIDs could be appealing also because they would provide higher assurance, but the lack of wide deployment in Europe prevents the adoption of this solution at the moment. It is, however, an area to monitor, and efforts are ongoing in both the AARC and the GN4 projects to follow the developments in eIDAS and the penetration of government IDs in EU Member States.
- Guest identities operated by, or for, research communities – these are identity providers set up by projects, infrastructures, research collaborations, or organisations supporting a specific research

community (as opposed to, say, restricting it to their own members of staff.) Typically, these are set up to provide a coherent "membership" service for people in the community, to grant them community-related authorisation attributes, to provide identities with a consistent set of attributes, or simply support users who do not have (or do not want to use) a suitable organisation to provide their identity. The community identity provider may have "grown" from a small collaboration with poorly defined processes, or with community-specific requirements in mind, and resources may need special adaptations to be able to use such identities. For example, it may be necessary to rename attributes to make them unique across the infrastructure, or to provide account-merging options for users who contribute to more than one research community.

- Commercially available services – it is possible to buy identities from commercial providers. Historically, commercial certification authorities issued certificates to users for (e.g.) signing email or signing code. By having users pay for their identity (as provided by the authority) – and pay for the associated validation – a provider could ensure that the identity meets a well-defined LoA. A more recent example is ORCID which provides identities for researchers to unambiguously link individuals to their work, and provides additional features for members (however, in this case members are organisations rather than individuals.)
- IdMaaS – Identity-Management-as-a-Service is a specific example of a commercial service. The classic use case is a smaller organisation that is unable or unwilling to run its own identity-provisioning infrastructure, and therefore sets up this type of service with a commercial cloud provider. As the services are run externally to the organisation, it could become easier to federate the identity, or to reuse it for single sign-on, at a later stage.

Users typically already use guest identities to access other services, so in this respect the use of guest identities is advantageous compared to using infrastructure-specific usernames/passwords: Users are less likely to forget them and share them. Moreover, policies can be defined that make guest identities still more attractive for reuse, for example by requiring that accounts not be shared, and the identity provider may monitor the use of identities to attempt to detect and limit misuse. Options also exist to enable the e-infrastructure to "strengthen" identities, by maintaining a reputation (rewarding users for positive contributions to the collaboration), or establishing peer-to-peer relations between users. There are many parallels for these types of interactions in the real world – after all, most human interactions rely on interpersonal relations and prior knowledge and reputations (as well as being subject to prejudices and biases) – but online these require more formal structuring as communications do not happen in person.

Any infrastructure can make use of guest identities, subject, if applicable, to legal constraints (e.g. permissions from the identity provider), technical constraints (use of protocols, interfaces, standards, and web/non-web access), and the appropriate levels of assurance for the rights granted in the infrastructure. It is recommended that a *risk assessment* be carried out based on the risks identified in this document and summarised in Appendix A. In terms of architecture, guest identities can be used directly by the service, or via a proxy (as described in the forthcoming MJRA1.4 document). When dealing with guest identities, the use of a proxy is recommended, because it allows communicating the LoA by adding it as an attribute to the metadata in the proxy, the guest identity providers need deal only with a single service (namely, the proxy), and the "strengthening" approaches mentioned above can be maintained via the proxy as user-level metadata. There are thus many options for the e-infrastructure to make use of – and manage the risks of using – guest identities. e-Infrastructures that serve multiple research areas should support identities with different LoAs, including guest identities, and many current e-Infrastructures already do. As the options to manage the risks increase in sophistication, the use of guest identities will increase further in the future.

1 Introduction

This document describes the main aspects to be considered when using guest identities in e-infrastructures. The recommendations focus on projects and e-infrastructures that *consume* guest identities, i.e. they provide resources to users who authenticate themselves through guest identities. In particular, recommendations also feed into the blueprint architecture document (MJRA1.4) due in M15 of the project.

It is foreseen that AARC will later publish an extended version of this document, which will provide more details and case studies, or alternatively split the information into separate, more focused, publications.

1.1 What Are Guest Identities?

The term "Guest Identities" generally speaking a loosely defined term that is commonly used within research collaborations to refer to any of the following:

- Identities for "homeless" users – users who are not otherwise associated with a participating organisation, including "long-tail" users. Similarly, in research collaborations community-based identities are commonly used to accommodate users who do not have, or cannot use, organisational identities.
- Ease of access – by making use of "lightweight" identities (that can be remembered by a browser, for example, without requiring regular re-authentication).
- Personal identities – where each identity is associated with a real person, and (by virtue of their original usage) less likely to be shared between individuals.

Note that the term "*identity*" is used here in the sense of an identifier or token presented to a service, typically as part of an authentication process, and representing the user in question to the service. Obviously, a user's real-life identity does not change, but in a distributed, online infrastructure, a user may choose different means by which to present themselves, and for data protection reasons may choose not to reveal all their attributes. It thus makes sense to talk about the user "having different identities" in this context. Moreover, we talk about *credentials* as identity combined with a "proof of possession" of the identity. The latter can be based on a shared secret or something more sophisticated, and the cryptographic strength thus forms part of the *level of assurance* (LoA) of the identity/credential. The link between the identity and the secret is established typically when the credential is created (and the identity is validated against the user's real-life identity); the proof subsequently only proves that the user is in possession of the secret: in other words, it only proves that it is the same user every time the credential is used. Authentication consists of presenting the identity and proving possession of the secret, thus proving that the authenticating user is the one to whom the identity was issued.

Guest identities can thus have the following characteristics:

- The e-infrastructure which relies on the identities has little control over the identities – their availability, the attributes published, the semantics of the attributes (when non-standard), and standards compliance. In particular there is little or no visibility of the processes behind the provisioning of identities (the processes are secret, undocumented, or can vary.)

- Guest identities are being used outside of the context for which they were originally intended, such as social media identities being reused within a research project, or a community identity intended for use by the community being used in an e-infrastructure that provides resources to the community.

- The connection between the guest identity provider (IdP) and the user may be looser than in the case of *"home" IdPs* (we use the term "home IdP" to denote the opposite of "guest IdP," i.e. a home IdP is an IdP run by the user's home organisation; see below for further discussion. By extension, "home identity" is an identity issued from a home IdP.)

- Some guest identities can offer improved privacy features; see [ENISA4]

The following definition for guest identities is assumed in this document, the rationale for which is explained below:

> *Guest Identities* are identities used outside of their original context.

Note that from the research infrastructure's perspective, guest identities could (almost) be defined as "anything but home organisation/federated identities." However, focus is given here to the *use* of the guest identity, as this is where the risks arise. Where identities are being used for their original purpose, it is expected that due processes and governance will be in place to ensure these identities are fit for purpose and meet all relevant requirements.

As an example, in the case of a research community that runs its own IdP, as long as this IdP is used exclusively by the community to access the community's own resources, there is no point in considering it a guest IdP. As it has been set up and operated by the community itself, it is assumed that ideally it has been designed it to meet its purpose, with an understanding of the risks. It is only when the community joins an e-infrastructure as a user community, and registers their IdP as an IdP for the e-infrastructure, that it becomes a guest IdP. The *risks are now transferred to the e-infrastructure*, so the e-infrastructure now needs to manage those risks arising from the reliance on the community IdP.

Clearly, there are infrastructure risks arising from the reliance on both home and guest IdPs, however, the purpose of this document is to discuss the risks specific to guest IdPs and how these may be managed.

In contrast to guest IdPs, *home IdPs* are operated within the context of specific organisations (Universities, Research Centres, Libraries, NRENs, etc.) and typically participate in national identity and access management federations. Typical traits of a home IdP can include:

- Well-defined and documented policies and processes, particularly policies complying with requirements of the target domain(s).
- Processes for auditing, e.g. on request or regularly, possibly by a third party (ideally an independent auditor.)
- Established and documented processes that are agreed with the federation(s) of which the organisation is a member. In particular, there is likely to be a documented relationship between the user and the organisation asserting the user's identity, e.g. as a member of staff, a student, or an alumnus. Moreover, this association can be communicated to the relying e-infrastructure.
- Additional features, such as traceability (proof of the association between the identity and the user's real-life identity) or monitoring (continuously validating the use of the credential with the aim of detecting and containing misuse).
- A legal foundation, such as being run by an organisation that is a legal entity.
- The IdP may have been designed with the needs of one or more specific user communities, or one or more (typically national) infrastructures in mind; the attributes it releases are tailored to the identified target audiences.
- A long-term plan – the hosting organisation commits to running the IdP over a "longer" period of time.
- Users generally cannot create their own identity (such as choose the names by which they are presented, or choose to be represented pseudonymously) or modify their own attributes; the identity is created for them based on an organisational identity management system.

[ENISA1] makes the distinction between execution (authenticating, establishing trusted links), management (creating and maintaining identity accounts), and governance. In comparing home and guest identities, it is notable that the execution layers are identical – being based, ideally, on standard and interoperable protocols, and some of the execution layers do of course need to be integrated with the e-infrastructure. Management is devolved for both guest and home identities. Indeed, from the perspective of the e-infrastructure, this is arguably the whole purpose of federated identity management (and AARC), to bring the management of credentials out of the e-infrastructure. Governance for home identities often lies within the national federations that the home organisations join, so is more closely aligned with relying e-infrastructures (which also become members of the national federations, as service providers). Thus, the relying e-infrastructures have, by virtue of their membership, some influence in the federations. In contrast, governance of guest identities lies in general wholly outside the e-infrastructures.

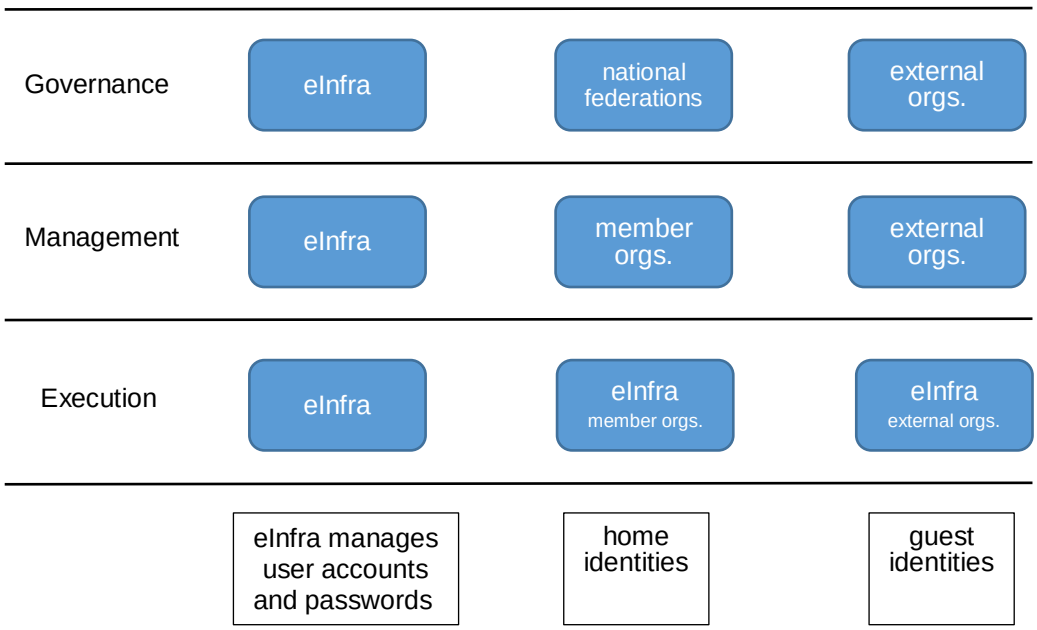

Figure 1: Comparing passwords, home identities and guest identities

Figure 1 compares an e-infrastructure managing usernames/passwords (the traditional approach, where all account management is done within the services) with home and guest identities. It can be seen that for the latter two approaches, there is much less burden on the e-infrastructure, as they need to deal with less of the stack. Moreover, in the case of federated identities, the governance typically lies with national identity management federations and their member organisations – these parties are generally outside the customer e-infrastructure that is relying on the identities, but still exist to (typically) serve academic and research communities. In contrast, the governance for guest identities lies with external organisations whose primary purpose is not to support use of e-infrastructures.

Note also that to assess the Level of Assurance (LoA) of an IdP, in general information is needed about all three layers. The execution layer is crucial for the LoA because it determines cryptographic strengths, security of protocols and implementations. The management layer is crucial for the LoA because it determines how users get their credentials in the first place, whether identities can be reused, the trustworthiness of attributes, etc. Finally, governance determines the target communities, the target LoAs, describing processes and requirements for the management as well as how to deal with incidents. Clearly, the more information the relying e-infrastructure has about each layer, the better it can make an assessment regarding the overall LoA to assign to the identities. (Note that it is always the relying e-infrastructure that should make the assessment of the LoA, as it is by definition the party relying on the credentials and needs to manage the risks. An IdP can assert its LoA based on established requirements and certifications, but this ultimately only serves to aid the relying e-infrastructure in its assessment.)

It follows that it is generally possible to assess the level of assurance (LoA) of home IdPs, whereas it can be more difficult with guest identities, thus introducing risks for the relying e-infrastructure.

Guest identities often lack documented processes necessary to establish their level of assurance, or do not publish this documentation.

However, the distinction between home and guest identities should not be considered to be a strict one, since some of the discussion in this document can apply equally to home IdPs, particularly when they are being reused in contexts outside of those for which they were set up (when they thus become "guest identities.") For example, an organisation running a home IdP within a national federation may one day be required to support a project requiring identities with a higher level of assurance (LoA). While this does not involve the use of a guest identity, some methods by which the LoA can be raised within the e-infrastructure are analogous to those that might be applied to a social media IdP.

This document looks at all the cases where guest identities can be useful and provides guidance on their use.

1.2 Why Use Guest Identities?

There are four basic reasons why a service provider might use guest identities:

1. Nothing else available – The home organisation of a user in a collaboration has not yet set up an IdP, or connected it to the national identity federation (and via the national federation, the eduGAIN interfederation.) Or the home organisation is in a country where the national identity federation has not yet been established, or has not yet joined eduGAIN.
2. Lack of coverage or harmonisation – Collaborations or communities may not be fully covered by existing federations. Users who are not covered will need a means of authenticating themselves so they can collaborate with users who are covered. It may also be the case that attributes are missing, e.g. if the infrastructure requires an email address and not all IdPs publish it, a guest IdP that does publish a usable email address is used instead.
3. Advantages of using guest identities outweigh disadvantages – For example, if users already use collaborative SaaS services in a public cloud which is linked to a guest identity platform (such as shared documents in Azure or Sharepoint, or Google docs/blogs), and the infrastructure forms a natural extension to this work.
4. Users only have, or are required to use, guest identities – An example would be where a community already has its own IdP (separate from national federations or other publicly available IdPs); membership of the collaboration could even be defined by their ability to authenticate to their IdP. ORCID is another example: while its purpose is not to authenticate users, it does issue and track user identities, and the infrastructure could require the use of ORCID in order to track publications of the work done in the e-infrastructure.

It is expected that scenario 1 will, eventually, become less of a problem. Here, it is primarily the home organisations themselves, aided by National Identity Federations, that should take initiative to deploy and connect their IdP to eduGAIN (in turn, the initiative should be prompted by user requirements.) Various NREN Federations have deployed services to support small institutions to easily deploy an IdP; an example is the 'IdP

as a service' offered by GARR in Italy. In addition, progress may be stimulated by a sufficient number of “interesting” services being enabled via eduGAIN. Moreover, through a series of projects, GÉANT has orchestrated several outreach activities that have paved the way and provided support for the setup of new identity federations in Latin America, Africa and Eastern Europe. This will incentivise institutions to set up their own IdPs, thus reducing the need for guest identities for users based at participating organisations in those countries. Obviously, if users use guest identities, they may need to migrate to institutional identities once their organisation gets a home IdP.

2 Options for Using Guest Identities

2.1 Managing Guest Identities

When users authenticate to an infrastructure using guest identities, options are needed to manage the accounts. These options may also apply to home (as opposed to guest) identities but are particularly relevant in the context of guest identities.

2.1.1 Level of Assurance

The processes followed by the guest IdP to validate the identity of the users may be unknown: they may be undocumented, they may be inconsistent across accounts (e.g. processes are changed, but the new processes are applied only to new users), or they may be documented but not published. Thus, the information needed to establish the LoA of the guest IdP is not necessarily available.

Erring on the side of caution, an estimated lowest bound LoA could be assigned to the IdP within the infrastructure . However, it should be remembered that the LoA of the IdP is not the complete picture. There are other compensating factors – discussed in the following sections – that can be implemented, and in practice, a careful management of risks is required (see section 3 for a discussion of risks).

2.1.2 Reputation

Reputation is an option that research communities have been, and are, investigating for enhancing existing credentials. In its basic form, it is a value assigned to the user (or to be precise, to the user's account) by the infrastructure. It can be argued that a ban of a user conveys the notion of a low reputation of the user's account, and, conversely, assigning authorisations and privileges to the account is an expression of, or at least presupposes, a higher reputation. In general, however, an implementation of reputation would be more fine-grained, with a transparent feedback to the user, so they know what their reputation is and which actions can improve it.

Note that the reputation is assigned to the user's account; so if the user's account is compromised, it is likely it will accrue a lower reputation – at least if it is misused. Conversely, a user may have multiple accounts with different reputation levels.

A basic example would be that of a user connecting with a guest identity account to an infrastructure. The identity is otherwise unverified (i.e. not checked by the e-infrastructure), but the user uses the same account every time. This is the case for "citizen science," where it does not really matter who the users are – anyone can sign up and use the service – what does matter is what they do when they are logged in. There are many projects and community sites where people may register and start building a reputation, often accruing points

either automatically or based on the feedback from peers. Some examples of such communities are Zooniverse, StackOverflow and Academia.

Some notable features of such communities are:

- Everyone can register – social media identities are accepted (and preferred), or failing that, people can register with an infrastructure-specific username and password.
- The reputation is assigned to the account. It is updated by monitoring the user's actions, or based on feedback from moderators or the user's peers.
- A given user's reputation is *visible* also to their peers. Users may thus be incentivised to increase their reputation, and there may be "ranks" they can reach, thus motivating them further to improve their standing.
- Users can re-register: thus, a malicious user may bin their low reputation account and start over.
- Users may have multiple accounts, each with its own reputation (which creates a risk if a user's reputation can be increased by their peers, the so-called "sock puppet" scenario, discussed in 2.1.3 and 2.2.1.)

Reputation systems are very widely used in general online services with user accounts and as such are quite mature, but there is no standard measure for reputation, nor is there a generally reusable framework or implementation. As an example, Confluence from Atlassian provides several flavours of reputation systems: the best "crowdsourced" answers "rise through the top through voting" [Confluence] and a peer-based reputational system: "You'll be surprised at how people will participate to earn points and climb the leaderboard. Top contributors achieve expert status." [*ibidem*]

It is possible that transferable reputation – within academic communities – would be an interesting topic in the future, e.g. within AARC2. A user joining a new community might use their existing reputation in another community as a basis for joining the new one – they would not have to "prove themselves" from scratch. This is of course analogous to real-world scenarios, e.g. where a prospective employee brings references from previous employers, or where an academic's reputation is based on their publications.

2.1.3 Using Peer-to-Peer Networks

Peer-to-peer networks can be seen as an extension of reputation management, as discussed in section 2.1.2, where, instead of assigning points by algorithms defined by the infrastructure, a user's "reputation" is based on their networks. This peer-to-peer network can be specifically targeted at the desired collaboration or it can be reused if it is available – a guest network based on guest identities. In its simplest form, the relation between two peers would be "is connected to," or "knows," or "trusts." Individual members could be scored on how much they are trusted by their peers – their "reputation" – or the connections themselves could be used, e.g. to set file sharing defaults. This, too, could be the topic of future research.

To some extent, the distinction between reputation as managed by the infrastructure (as discussed in the previous section) and reputation as managed by peers is quite a fine one. In the former, it is usually the infrastructure itself that defines rewardable actions, and/or reputation is managed by trusted moderators; in the latter, anyone can in principle vote for anyone else. There are thus additional risks arising, whether the

peers can be trusted or they collude, or in the worst case they all represent the same person ("sock puppets"), when a user has multiple accounts and pretends they are different people.

A network-related example is the WebID proposal from W3C, in which each user can publish their list of "friends" – essentially leading to a graph database of users[1]. If the users already maintain a network of their collaborators, this network could be used to manage accounts and authorisations, provided the infrastructure allocates resources to the collaboration as a whole and not to the individual.

The IGTF [IGTF] has a PGP-based network to enable secure communications (digitally signed and optionally encrypted) between participants (i.e. between operators of national CAs and representatives of the larger RPs). As with any PGP network, keys were verified via their fingerprints at face-to-face meetings. While this network uses an established PGP infrastructure, its reusability for "guest" (authorisation) purposes would require querying the relevant key server for peer-to-peer signatures and building a representation of the whole collaboration.

In a rudimentary form, this peer-to-peer model is widely used in research infrastructures: users are networked by virtue of being members of the same community, resources are allocated to the community, and the community is trusted to figure out how to use the resources.

More generally, is there scope for building academic peer-to-peer networks and reuse them across infrastructures? People have their social media IDs and use those to connect to their peers, but these networks are not accessible to the infrastructure, even if the identities are. It seems unlikely they want to build and maintain yet another social network – an infrastructure-specific one; it is not clear even whether users will be inclined to vote on each other's contributions or "like" each other's pages, particularly as there is no benefit for them. (Recent work, however, adds "gamification" to encourage users to participate; this is provided as (commercial) third-party add-ins to (commercial) collaboration platforms.)

It seems at this point more likely that an infrastructure seeking to make actual use of its users' networks needs to (a) either obtain this network from somewhere where it is already established and maintained, or (b) mine the user interactions in the infrastructure, using data mining or similar techniques. This, too, could be the subject of future research.

2.1.4 Using Supplementary Information

Additional information could be provided by other authorities. A user may register with an e-infrastructure using a guest identity, but the guest IdP may have a low LoA, or certain attributes may be required but not supplied by the IdP. In this case, it would be natural to ask the user to supply other *trusted* sources of this information, provided these additional sources are independent of each other (and of the original IdP).

The simple use case is that the IdP provides an identity for the user but does not know the user's role in their collaborations, so the user's collaboration supplies an *attribute authority* which can provide additional attributes about the user (and the user chooses whether or not to assert these attributes.) One can extend such a scheme to provide additional features: users can link different IdPs to their infrastructure account, or trusted attribute authorities could be queried as to whether they have any information about a given user.

[1] Mike Jones from the University of Manchester proposed such a scheme [MJones]

One of the most widely used cases today is where VOMS as an attribute authority is used with X.509 certificates from [IGTF] as identities. While VOMS is based on RFC 3281 attribute certificates [MJRA1.1], another technical option is SAML.

The future ORCID functionality (Appendix B) is another example of providing additional information, where user-asserted information can be supported by their home organisations. This case, in particular, is expected to gain importance, as users will need to provide their ORCID IDs in order to track the associations between their data, their work on the infrastructure, and their publications, and to prove to the funding bodies or infrastructure accounting that they have done so.

Finally, two factor authentication and ELIXIR's "step up authentication" are also examples of using supplementary information to strengthen the original credential.

2.2 Types of Guest Identities

In this section, the various types of guest identities are briefly surveyed, as these will each have different advantages and disadvantages.

Note that identities for libraries were removed from this document. The project will publish information about libraries elsewhere; the reader is referred to [AARC] for further information.

2.2.1 Social Media

Social media identities are those from Google, Facebook, Linkedin, etc. Typically, these identities are available for authentication to resources, with no special agreements being needed with the IdP (i.e. one does not need to exchange IdP/SP metadata), but the underlying network of connections – "friends," "collaborators," "colleagues" – is not available.

2.2.1.1 *Advantages of using social media identities*

1. Most people have one or more already.
2. While anyone can create a Google/Facebook/etc. account, people who use theirs for more than throwaway interactions – if they use them to liaise with colleagues, school friends, family, etc. – will protect them better than they would an account created specifically for the interaction with the infrastructure.
3. The social media platform is likely to implement proactive monitoring of the accounts, as they have an interest in avoiding malicious users, or, if applicable, the sharing of accounts. While these processes are typically not fully explained to the service providers relying on their identities, it does at least improve security in a way that is not typically found in infrastructure-specific identities.
4. Social media, by their very business model, tend to be "nosy" about what their users are up to, so also have an interest in their identities being of a high enough quality to be reused, and to allow their reuse. The social media platforms thus implement monitoring of the use of the identity to reduce the

risk of abuse. (Arguably, this "nosiness" is also a disadvantage, for privacy reasons, see the following section.)

5. As the FIM4R activity pointed out [FIM4R], younger researchers are usually more comfortable with the use of social media identities for various activities of their daily life, and may expect to continue to use these identities for their research.
6. Linking social media identities with the work in the infrastructures may make it easier for users to disseminate or communicate about their work on social media, making it possible for infrastructures to link more closely with social media. For example, a user who is already on Google+ and using their Google ID to authenticate to their research infrastructure will also be able to use other Google services, such as blogs or email, as they need only log in once. (Again, there may be privacy issues here, as the blogs and other services will track the user's work.)
7. Social media identities are easy to integrate with the e-infrastructure: no special agreement is needed with the IdP as the users themselves consent to the attribute release.

2.2.1.2 *Disadvantages of using social media identities*

8. The SP has no influence over the IdP, so the IdP may change technology or policies without any consultation with the SP. It is not the IdP's main role to support the provisioning of identities outside of the social network.
9. There may be internal checking processes in the IdP but it is not always clear what they are. Indeed, the IdP may hide their internal processes to prevent malicious users from working around the checks.
10. Privacy. The provider of the Guest identity will/may be tracking the user's activities and correlate that data to other online activities the user engages in. This can happen directly if the user authenticates using a social media ID, or indirectly if the user is authenticated to the social media and the infrastructure happens to load something (such as a "like" or "share" button) from the social media platform. Indeed, some platforms have extensive means of monitoring users (cf. Google Analytics), and an all but limitless capacity for storing and mining this information.
11. Conversely, if the user links their social media ID with other ID(s) that are used for day-to-day business, loss of privacy can happen in the other direction: the other IdP or the infrastructure may gain attributes about the user from the social media platform. Users should be informed about the potential risks of linking their social media IDs with their work activities.
12. A user could create an account pretending to be someone else (i.e. in someone else's name), or could create fake accounts, either to circumvent restrictions placed on one account, or to create "sock puppets" (user accounts purporting to be collaborators but are really be the same user voicing support for themselves). A fake account could also be used as a basis for "social engineering" attacks (where for example someone is persuaded to release restricted information to an attacker pretending to be a person to whom the release should be authorised.)

2.2.2 Community Identities

Many projects, research communities, or other collaborations, have set up their own IdPs, in order to have a harmonised credential with a unified set of attributes, or to support users who do not have (or do not want to use) a home organisation.

Appendix B contains an overview of a representative sample of community identity providers, illustrating their (re)use in projects as well as for the community themselves.

2.2.2.1 *Advantages of using community identities*

13. The IdPs come with existing user communities. The e-infrastructure need only trust a single IdP to provide services to the whole community. Users are members of the community if and only if they can authenticate via the IdP.
14. Users already use their IdP with other community-related services, so know how to use it, and, presumably, have some interest in protecting the credential.
15. As users already use the ID with their existing community resources, it could make it easier to provide integrated services to the community, e.g. where users need to authenticate only once to their IdP. Users could simultaneously log into a community and an infrastructure portal with the same login.

2.2.2.2 *Disadvantages of using community identities*

16. These IdPs tend to live only as long as the project that funds them – running an IdP in production does need some resources, as does user support, so a sustainability plan may be required before the IdP can be used.
17. The level of assurance is often ill defined, as identities are typically granted lightly to people in the research community, or the community may have grown organically, or been established at the inception of a project. Even if there is a defined process, it is usually not rigorously documented, let alone documented in a format that aids the establishment of the LoA by someone external to the community.

2.2.3 Government and Bank eID

In July 2014, European co-legislators adopted the Regulation (EU) N°910/2014 on electronic identification and trust services for electronic transactions in the internal market (eIDAS Regulation) [X1]. The purpose of eIDAS is to ensure that people and businesses can use their own national electronic identification schemes (eIDs) to access public services in other EU countries where eIDs are available. The regulation effectively creates a European internal market for e-Trust Services, namely electronic signatures, electronic seals, time stamp, electronic delivery service and website authentication, and ensures that they will work across borders and have the same legal status as traditional paper-based processes.

An implementation timeline is already in place, starting with the adoption of the regulation in 2014, then continuing with the voluntary recognition of eIDs by the European Member States starting from September 2015. At the time of writing of this document (Q2, 2016), the Connecting Europe Facility (CEF) is making available the first version of the eID Interoperability Infrastructure. In parallel, the European Union co-finances 20 projects addressing the cross-border interconnection of eID services in Europe, in an effort to support the EU Member States with the rollout of technical infrastructure, aiming to create interoperable, pan-European eID services under the eIDAS legal framework. The next milestone in the timeline falls in July 2016, when the "Trust Rules" will become applicable and the Member States can voluntarily use the EU Trust Mark. The timeline ends in September 2018. At that point, the cross-border recognition of eID between the Member States should have already been established and the cross-border authentication using eIDs should be operational.

Given that in the next two years, citizens of most European countries will have access to national eIDs, it is logical to expect that research collaborations might become interested in leveraging the eIDs for user authentication. Indeed, the ability to have stronger bindings between the physical persons and their digital identities might provide a solution for security/privacy sensitive scientific domains. Furthermore, if the uptake of national eIDs is high (as expected), then suddenly a large portion of users in the research and educational space, especially those who are part of the so called "long tail of science", will have access to high quality eIDs, even if they are "citizen scientists" or affiliated with organisations without the proper infrastructure.

2.2.3.1 *Advantages of using government or banking eID*

18. They are/will be backed by EU member states, and will be used in a growing number of citizen transactions ranging from the interaction with state services to banking and health care. eIDs will be used wherever one would show passports or traditional IDs today.
19. National eIDs can have a high level of assurance [ENISA3], which is a requirement by those scientific communities dealing with sensitive private data. Indeed, as with recent European passports, physical eID tokens may well support some form of biometrics, which while not perfect have the potential to provide high LoA while not affecting usability much.
20. Citizens will be using them all the time (presumably) for services, so there will be significant support infrastructure for making their use ubiquitous.
21. As they are an important part of managing their (online) identity, people will keep them safe, and there is support for protecting the data on the card from unauthorised parties [ENISA4].
22. With a wide use, it is likely that enough effort will be put into supporting eIDs on or with mobile devices.

2.2.3.2 *Disadvantages of using government or banking eID*

23. Typically, national eIDs will come in the form of hardware tokens/smart cards holdings the citizen's key pair(s). Interfaces may be needed to interface to the cards with keys on them and, at least in the beginning, the learning curve will be steep.

24. Use of national IDs may be more restricted by policy; one option here is to restrict the national ID to an initial transaction (typically an initial registration) and use another credential for subsequent transactions with the infrastructure.
25. Users may not want to use their national IDs within e-infrastructures for fear of compromising the ID, or otherwise exposing it. The ID may be used with other resources which users perceive as more valuable (such as their personal bank account), cf. [ENISA3].
26. Users may worry about the release of attributes from eIDs as these are linked more closely to their "real-life" identities. In particular, using national eIDs might be a problem especially in cases where anonymity is important (where the research is controversial) – however, see also [ENISA4].
27. Cross-border authentication using national eIDS is ensured by the "eIDAS Interoperability Framework". Service Providers, who might want to make use of national eIDs for authentication of their users, will need to integrate their services with this framework. This might prove difficult if the technical requirements are very different from those that currently apply, or might even be prohibited due to policy/regulatory constraints as mentioned in the second point. See also [ENISA5].
28. The eIDAS regulation and the eIDAS Interoperability Framework is relevant for the European Member States. Research collaboration is global, so the use of national eIDs through eIDAS can only be a partial solution to the problem, at least for the foreseeable future.
29. Not all countries have them, or make them compulsory [ENISA4]. Indeed, in some countries (such as the United Kingdom) there is stronger cultural opposition to national identity cards than in others (such as Germany) [UKID].

2.3 Commercial Providers

2.3.1 Certification Authorities

Commercial certification authorities implement checks that secure services on the web; in the past they have also offered personal certificates of a high level of assurance, e.g. to sign email or code, to make digital or electronic signatures[2], or to authenticate users to public services. In this sense, this is an extension of government IDs based on X.509 certificates, but offered by a commercial provider. The advantage of certificates is that X.509 is widely supported as a technology, and certificates can provide excellent security if managed correctly. This is because:

30. The secret is not exposed upon authentication (zero-knowledge proof of possession).
31. The secret is (typically) two-factor, consisting of something the user possesses (an encrypted private key, either in a file or on a key token), and something they know (a passphrase, resp. pin, to unlock the key/token).
32. They can use strong cryptographic security, including hardware with certifications such as NIST FIPS140-2 [NIST].

[2] The difference is whether the user signs automatically, e.g. via their mail user agent, or signs intentionally, as if signing an agreement.

The main disadvantage is that many – perhaps less technical – users find them hard to use without the proper tools, and browsers (as well as e-infrastructures, [GFD.225]) have proved very inconsistent in how they manage certificates.

2.3.2 IdMaaS – Identity Management as a Service

IdMaaS is a type of commercially offered service where some or all of an organisation's identity management is implemented "in the cloud."

Advantages of this approach include the usual cloud ones – that one need not have staff to run the service (except for user support), it is possible to have high availability setups, and commercial cloud data centres are run to an exceptionally high standard. A more specific advantage is that cloud providers can offer sophisticated monitoring to catch misuse, or attempts to compromise the service. Another advantage is that it may be easier to federate the identity provider because it already has authentication endpoints outside of the organisation. Disadvantages are also the usual cloud ones, the running cost, the fact that networks must be available to connect to the service, and that organisation-critical data must be entrusted to a third party.

An IdMaaS can also support multiple protocols, and can therefore synchronise internally with an existing on-premise identity management system (typically Microsoft Active Directory) but also with external ones via, e.g., SAML or (SOAP) web services. Examples of this include OneLogin and Microsoft's Azure AD.

See also section 2.4.5 for a discussion of the academic side of IdMaaS.

2.4 Deployment and Architecture Considerations

This section describes some specific aspects to be considered prior to deploying guest identity services for infrastructures or communities.

Briefly, a community that wishes to obtain identities for its users has the following options:

- Set up their own IdP, run by one of the participant organisations or in the cloud (see IdMaaS in sections 2.3.2 and 2.4.5), and start registering users.
  - Advantages: full control over IdP. Simplified attribute management – authentication is equivalent to membership;
  - Disadvantages: need to maintain the IdP throughout the lifetime of the community; need to register all users from scratch; need to consider survival beyond the end of the project. Potential proliferation of IdPs in infrastructures (section 2.4.1).
- Use someone else's identity provider (section 2.4.2).
  - Advantages: someone is already running the service; users may already be registered and may be using the identities for other things. Fewer IdPs in infrastructures.
  - Disadvantages: less control over attributes released. Need to agree who will be doing user support, as the existing IdP may not be willing or able to support a new community.

It should be clear that it is better to use someone else's identity provider, if possible. This is no surprise: most e-infrastructures and communities wish to focus on their target research, not identity management – hence also the need for AARC.

Fewer IdPs also means a lesser burden on the relying e-infrastructure, as it needs to configure and liaise with fewer identity and metadata providers. In practice, establishing the trust link between an IdP and an e-infrastructure (or proxy) and getting the identities working takes some time (a few days, in the best case, weeks or months in the worst cases). Using the layers introduced by Figure 1, it requires effort both in the execution layer (exchanging metadata, agreeing attributes, technically connecting the two) and in the governance layer.

2.4.1 Scalability

As mentioned in the previous section, connecting an IdP to an e-infrastructure (or proxy) is not trivial, and can thus cause scalability problems. Moreover, more IdPs also means longer WAYF-lists (Where Are You From, the process where the user selects their IdP); while incremental search is quite efficient and widely used, unintentional scalability problems can arise. For example, the EUDAT B2ACCESS service displayed the icons of each home IdP in the selection list, loaded from each remote IdP as directed through the IdP metadata, thus slowing down browsing of the list (and also leaking information to the IdPs).

We note that some types of guest IdPs can help address scalability issues by reducing the number of IdPs required. Social media IdPs and government eIDs can potentially authenticate millions of users, whereas community IdPs can make the scalability problem worse if they support only tens to a few hundred users.

2.4.2 Reuse of Existing IdPs

A community may wish to use existing IdPs for its users, to authenticate them to its own services and/or to e-infrastructures that make resources available to the community. In this scenario it would clearly be best to prefer suitable IdPs where users are already registered, in order not to have users re-register and maintain yet another identity. If there is more than one remote user IdP, a proxy scenario is preferred (to be described in the future milestone document MJRA1.4 [AARC]) – the proxy then becomes the IdP for the services.

The following questions then need to be considered:

- Do the identities have a sufficient LoA? – can the risks (section 3) associated with using them be managed? This includes sustainability; it may be wise to have a plan for transfer of identities in case of termination of the IdP.
- Can the identities be reused? – if the proxy has signed up to the eduGAIN data protection Code of Conduct (CoC, sometimes known as CoCo), some thought may have to be given as to whether attributes can be published to the community's services (which may become "third parties," and there may be services outside the European Economic Area.)
- In particular, are the attributes published by the IdP adequate and relevant for the community services? – For the IdP, there is an added risk as they are now sending their attributes to additional

services, so some sort of assurance regarding the use of attributes by the community services may be needed.

2.4.3 Usability

Social media identities have excellent usability when it comes to web-based services. As people access social media platforms via web browsers, they tend to remain logged in, and they then only need to authorise the release of attributes to make use of this identity.

A related question is whether social media identities can be reused as easily on mobile devices; generally social media platforms have their own apps and a custom app will need to integrate into its identity management framework.

Another related question is whether guest identities can be used for non-web services. If the use of the guest identity is primarily web- and app-based, as is the case with social media and eIDs, one may need a translation or conversion to a separate identity which can be used e.g. on the command line. See e.g. [MJRA1.1], section 2.2.

2.4.4 Proxy

When deploying guest identities, it may be useful to use a proxy (see MJRA1.4, [AARC]) to manage the authentication. First, in terms of administration, only a single service is registered with the guest identities; secondly, translation and harmonisation may be needed if the guest identities are diverse (or are used in combination with home identities); but more importantly, supplementary attributes can be added by the proxy, specifically to manage the diversity of the IdPs; see sections 2.1.2 and 2.1.4.

2.4.5 IdMaaS – Identity Management as a Service

One specific option for research communities or e-Infrastructures is to set up an IdP as a cloud-hosted service. It can be set up as a primary IdP – it allows research communities to manage user accounts with its own database (green components of Figure 2). It is possible to extend this to a proxy service (blue elements in the Figure). In fact, the proxy IdPs elements that are being deployed in Research Infrastructures and e-Infrastructures for the research communities can be considered as a special kind of IdMaaS, which can connect the users' "home" IdPs to the infrastructure and offer further Identity Management capabilities on top.

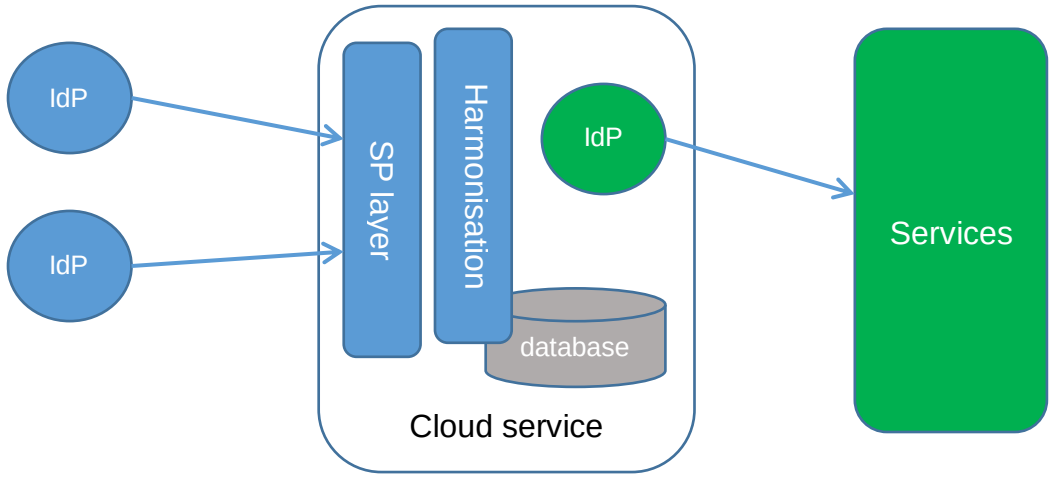

Figure 2: IdP or Proxy in the cloud

The advantage of setting up in a cloud environment, apart from the usual cloud characteristics of availability and pay-per-use, is that:

- An IdMaaS is a hosted service, so is easier for the site to set up and run; they require less expertise and fewer (technical) operational interventions;
- A proper IdMaaS solution offers additional security monitoring or other protections, e.g. monitoring where users are logging in from, whether there are too many failed logins, etc.

Microsoft Azure AD is an IdMaaS service integrating site Active Directory (AD) services via an IdP running in Azure. This IdP can provide its own identities, can use the site AD identities, and can also integrate fully or partially with the site AD, so enables users to reset their passwords, or administrators to manage attributes. A similar solution is the "IdP in the Cloud" service provided by GARR for academic institutes in Italy.

The EGI AAI Platform and EUDAT's B2ACCESS service are potential solutions which can enable e-Infrastructures to provide IdMaaS capabilities for research communities. See also section 2.4.2.

A discussion of the commercial services for IdMaaS can also be found in section 2.3.2. (Using a public cloud is clearly always a commercial transaction; the main difference between section 2.3.2 and this section, however, is whether the IdP is wholly commercial/outsourced.)

3 Risk Management

In any project or infrastructure, the reliance on external components introduces risks, and an infrastructure's reliance on external IdPs (whether guest or not) is no different. Using guest IDs, however, may introduce specific risks to the e-infrastructure, which may or may not be as prevalent with home IDs.

The following section provides a *brief* summary of these risks, which integrated with work from AARC NA3 on LoA and with case studies of the use of guest identities in current e-infrastructures should provide a full risk management framework. See also Appendix A where these risks are summarised in tabular form for convenience.

- *Policy of (re)use of the guest identity* – Is the infrastructure allowed to use the guest identity? Are there specific constraints that need to be respected?
- *Documentation regarding guest identity processes not available* – e.g. it is proprietary, or was never written down, or the processes are documented but the participants do not follow them – see section 2.1.1.
- *Lifetime of guest IdP* – as the guest IdP is external to the infrastructure, its lifetime may be limited, particularly if it is run for a community by a project whose funding will come to an end. It becomes necessary to consider how to migrate users off the IdP towards the end of its life. In contrast, home IdPs tend to have a longer lifetime, or at least the assurance of a longer lifetime.
- *Initial signup, role of (human) administrator vs "cloud" registration* – How are users initially registered with the service? If registration needs to be approved by an authorised person, the information that is available to the authoriser from the IdP, and any supporting information (section 2.1.4), needs to be considered. If on the other hand registration is automatic (a "cloud" type registration, *i.e.* without the intervention of a human administrator), it should be considered whether policies connected to the IdP can help protect the infrastructure against misuse.
- *Additional information - e.g. dual roles, two-factor login where users add a mobile number or a credit card number* – The user can be asked to supply additional information (section 2.1.4), or supporting information can be maintained by the infrastructure (section 2.1), in order to "strengthen" the guest identity and/or help manage risks (such as any arising via the disadvantages identified in section 2.2).
- *Incident handling* – If it is necessary to be able to ban a user (for misuse), it should also be possible to ensure that users cannot just create a new account and continue their misuse. Moreover, where a user has used the infrastructure inappropriately, *traceability* becomes important. There should be a means to link the attribute to a real-life identity, in a form that could be presented as evidence in a court of law in a worst-case scenario. Of course, these issues are equally valid for home IdPs, which however are more likely to have already addressed them.
- *Scalability* – Some guest identities are designed to support a specific number of users, e.g. how many they can comfortably accommodate in their databases, or how many can simultaneously log in. Facebook has more than 1 billion active users, but smaller community IdPs may not appreciate the additional logins generated by their IdP being reused. Conversely, as linking IdPs to anything takes

time – particularly if work is needed at governance level (Figure 1), adding more IdPs becomes a burden on the infrastructure.

- *Compliance with Network and Infrastructure AUPs* – Any information carried by networks has to comply with the law of the relevant country, or countries, as well as the terms and conditions of the provider: NRENs always have AUPs. When using a guest identity outside of its original geographic or network domain, its use may span additional NRENs (or countries) whose AUPs/laws were not originally considered. The infrastructure may need to seek additional user consent for compliance with the additional AUPs (see also the following point).
- *Using Guest Identities: Authentication and Signatures* – For the most part, authentication is considered the basic use of guest identities. Authorisation has been briefly discussed in sections 2.1.2, 2.1.3, and 2.1.4 (see also the following point.) However, additional uses are sometimes required, e.g. signatures to confirm acceptance of AUPs (see previous point). Traditionally, digital signatures are made with X.509 certificates, whereas guest identities rarely use X.509, and many of the user communities would be unable or unwilling to handle X.509 (let alone be able to digitally sign a document.) While this topic is too complex to be covered extensively in this document, it should be considered whether a signature can be adequately recorded and protected against user repudiation.
- *Assigning Authorisation to Guest Identities* – When authorisations are assigned to any identity, it should be considered whether the information presented (typically a request for resources in some form) can be trusted – is the authorisation granted to the right user? Is the presented information current? Does the LoA (and supporting information, section 2.1) support the increased privileges? For guest identities, the disadvantages (section 2.2) could mean that these questions require more careful consideration or additional risk management.
- *Attributes and personal metadata* – By definition, guest identities are used outside of their original context, which means that the attributes available will in general not be tailored to their new use. It should thus be considered whether the right attributes are available – also noting that attributes that are not needed should not be requested (cf. the eduGAIN CoC.) Moreover, as the guest identity is used for other purposes, it may become possible to infringe on the user's privacy by comparing their use within the infrastructure with their use of the same identity elsewhere. This "collusion" between service providers is sometimes a feature – users are presented with a consistent environment across providers – though caution should be used as unintended consequences could arise. (See also the following point, and [ENISA5] for further discussion of the risks associated with using eIDs in particular for cross-border authentication.)
- It should be noted that attributes can come from different sources – from the IdPs, from attribute authorities (section 2.1.4), from the user's peers (2.1.3), or be maintained within the infrastructure itself (2.1.1, 2.1.2). Not all of these will have the same level of assurance or trustworthiness. It follows that the level of assurance or trustworthiness assigned to the user's profile as a whole will either need to follow the baseline, to be that of the lowest level, or will potentially need to be maintained for each attribute. This results in a classification similar to an attribute-level LoA, a concept which has been used generally for email for example (verified - not verified), and is also emerging in infrastructure projects such as EUDAT that make use of attributes from several sources (where a balance needs to be struck to keep such an LoA sufficiently simple.)
- *Data protection concerns (personal data)* – data protection guidance aims to give users control over their use of data. Some IdPs let users control which attributes are presented to the SP, but a fine-grained attribute release is not always technically supported or practically possible.

- Furthermore, with reuse of identities between services attention should be paid to the "collusion" between services (see previous point): if adequate and relevant attributes are released in a targeted way, it might be possible to gather more information about the user by combining attributes from different services. With the GDPR, it is likely that further guidance is needed both for the users and the infrastructure providers; it seems likely that additional technical features will be needed (e.g. a means for users to see "how were my attributes used in the infrastructure?"). See also the related *Privacy* below.
- *Schemas and Attributes and Interoperation* – As regards reusing attributes outside of their usual context, the issues of trustworthiness (e.g. in section 2.1) and of which attributes are available (see Appendix B) have been discussed. However, technical interoperations are also needed if attributes are used directly (*i.e.* at the execution level in Figure 1), as opposed to used via a translation service. Are the attributes published using the same schema (such as eduPerson), and are they published with the same semantics? For example, the eduPersonScopedAffiliation has a set of recommended attribute values (member, staff, alumni, etc.) which have been subject to varying interpretations.
- *Privacy* – as discussed briefly in section 2.2.1.2, privacy can be a concern when giving users access to (remote) resources, a concern that needs to be weighed against the need to protect the infrastructure against misuse, and accounting for (appropriate) use. Infrastructures emphasising privacy have allowed the use of attributes for presentation (i.e. maintaining the user's account across logins) and statistical purposes; any additional use (e.g. authorisation, collaborations, etc.) has required user consent. As users leave digital trails in their research as well as in their personal (online) lives – see also the related *Data Protection* above – it may be useful to enable users to define their privacy levels. In their real (as opposed to online) lives, people tend to be less private with their family and friends, and more private with complete strangers, and perhaps somewhere in between with their colleagues and collaborators. As private and working lives can overlap – *e.g.* when social media identities are used also for research – more research should go into enabling users to define their boundaries.

4 Conclusions and Future Work

The use of guest identities in infrastructures is a requirement to support "long tail" research and user communities that already have community-specific identities. Many infrastructures already offer support for guest access; for social media it may link a social media platform to a user's work platform, and for community identities it makes it easier for the infrastructure to support communities. Indeed, communities can themselves easily set up their own identity provider, although this may lead to a proliferation of IdPs and, worse, to users having separate identities for each community in which they participate.

Naturally, the guest identity approach has advantages and disadvantages, both from the user and the infrastructure perspective. Additional risks arise, or become more acute compared to risks from home identities such as those from NREN federations. These risks need to be managed. In section 3, the main risks have been outlined. These risks are different for the various types of guest identities, but generally lack documentation, or at least documentation that is publicly available. Risk mitigation spans a range of topics. It can be technical (implementing means to control attributes, or to see how they were processed) or procedural (adding supporting information). Mitigation may be related to human resources (training, awareness, etc.), or can involve legal issues (imposing additional AUPs), or include information from collaboration (reputation, making use of peer networks). Many of the proposals involve supplementary attributes, whether from external sources (peers, attribute providers, two-factor authentication), or internally (LoA, reputation), where it should be considered whether the additional information is trustworthy and independent.

As regards the infrastructure, connecting guest identities is technically relatively easy, due to the wide use of standard formats and protocols such as the SAML protocols, OpenID Connect, and X.509 certificates. Managing identities via a proxy can also help control the risks arising from the use of guest identities.

It is clear that a multi-LoA approach is needed for authentication. Moreover, as attributes can be combined from different sources, potentially LoAs need to be maintained at the attribute level: this would enable the infrastructure to authorise users to access "expensive" resources using very trustworthy attributes even if the user authenticates with a lower LoA resource (as long as there is little risk of the account being compromised.) Indeed, such work is already being proposed in EUDAT. This approach could increase the usefulness of lower-LoA IdPs or attribute providers, although care should also be taken not to over-complicate attribute management.

4.1 Future Work

This document has described a number of areas that may be of interest for future work, some of which will be started in the remaining year of the AARC project, and may be carried out within or in collaboration with the e-infrastructures. These areas are listed below, starting with the basic future work and going on to developments to be investigated further.

- The risks need to be expanded into a full risk management framework.
- The risk management could be aided by case studies in the use of guest identities, particularly focusing on how users sign up (and sign AUPs), and how they obtain authorisations. It would also be interesting to investigate whether e-infrastructures actually can make use of social media and ORCID identifiers to link users' work in the e-infrastructure to their social media platform, and to their publications, respectively.
- Migration of identities – from one identity to another – for a larger group of users would also be an interesting topic for a case study. In particular, this can feed into managing the IdP lifetime risk.
- Making attribute authorities available to communities, based on existing identities with well-defined levels of assurance, whether guest or established.
- In particular, the use of government eIDs as a high assurance identity for e-infrastructures needs to be explored further.
- Additional data protection guidance will be needed for the use of guest identities; [ENISA3] and [ENISA4] looked at eIDs but more work is needed for social media; also user control/consent and perhaps visibility of how their attributes were used.
- How are non-web and mobile access to be supported?
- Can reputation systems usefully be put into practice in e-infrastructures?
- It may be interesting to look at reputation that is transferable between e-infrastructures and/or communities (section 2.1.2).
- Could useful networks between researchers be derived, akin to social media networks? (section 2.1.3).

Appendix A – Risks

The table below summarises the risks from section 3; it will form a basis for a more elaborate risk management framework in the second year of the AARC project (by combining it, for example, with work from AARC NA3 on LoA, as well as with case studies from the e-infrastructures that currently use guest identities).

The "owner," *i.e.* the stakeholder that principally needs to deal with the risk, or at least has an interest in the risk being managed, is shown for each risk. It can be seen that the e-infrastructure owns most of the risks; for some, ownership is shared with the community – the exact ownership can depend on the exact nature of the risk. For example, if the right attribute is not available from the IdP (number 11 in the list), it can impact the community (they need to provide additional attributes) or the e-infrastructure (relying on attributes which are not precisely right.) The table should serve mainly as an indicator of longer, more precise, work, to be published later in the project.

Risk	Main owner(s)	Mitigation
1. Policy/governance – can the guest identity be reused, are there constraints that need to be satisfied?	e-infrastructure community	Seek guidance, e.g. from ENISA (or AARC), based on experiences (e.g. Appendix B)
2. Documentation of identity management processes not available (does not exist, or is not published), or participants do not follow it	e-infrastructure	Section 2.1
3. Lifetime of guest IdP?	Community	Need a migration/exit strategy; supplement IDs with persistent identifiers
4. Initial sign-up with guest IdP? How are users registered with the guest IdP, and, subsequently, with the service?	e-infrastructure	Delegate to communities or principal investigators (and allocate resources to them). User credit card or user/organisational subscription (needs accounting).
5. Additional information – if users or their community need to supply additional information, it may come with a different LoA from the IdP	e-infrastructure	Section 2.1.1
6. Incident handling – will the guest IdP cooperate	e-infrastructure	Develop mitigation planning using

Risk	Main owner(s)	Mitigation
in the case of an incident?		input from NA3
7. Scalability – are there scalability issues arising from (re)using guest identities?	e-infrastructure	Base expectations on experiences in e-infrastructures and/or SA1.
8. Compliance – are there any issues with compliance with network and infrastructure AUPs when guest (as opposed to home) identities are used?	e-infrastructure community	Seek guidance e.g. from/with GEANT or eduGAIN as appropriate.
9. Are all uses of identities covered – guest identities are normally used for authentication only	Community	Additional infrastructure may be needed. SA1 work and work in infrastructures on attribute authorities can help.
10. Are there issues arising from assigning authorisation to guest identities?	Community e-infrastructure	Complex risks, depends on use in e-infrastructure – accounting could be repudiated by user, or assigned to wrong user
11. Attributes and personal metadata – risks of inappropriate use of attributes, or the right attributes are not available	Community e-infrastructure User	Complex risk: needs careful assessment probably on a case-by-case basis based on IdP documentation; case studies would help (and compliance to standard schemata and semantics)
12. Compliance risks with data protection – using guest identities leads to further places where its attributes are used	Community e-infrastructure	Complex risk (with GDPR); data minimisation helps; seek (and manage) user consent
13. Are there technical risks, e.g. in protocols and schemata being governed by external organisations?	e-infrastructure Community	A proxy can help manage this risk (c.f. 2.4.4), as changes can be "absorbed" by the proxy.
14. Privacy – are there privacy risks for the users using guest identities?	User (data subject)	Help user understand privacy risks – guidance, training.

Table 1: Risks

Appendix B – Examples of Community Identities

This subsection provides several examples to illustrate how project-specific or community-specific IdPs are used in academic environments. It is not meant to provide an exhaustive list. Furthermore, some examples are not "academic communities" in the strict sense of the term. No comparison of these communities is attempted here; the descriptions are included to serve as illustrations or examples of the points discussed in the main text.

CILogon

CILogon is a US project providing both a service, and a software stack to enable federated access to an online IGTF-approved CA: users authenticate with trusted SAML or OpenID-Connect identities and obtain X.509 certificates. The software stack is based on a MyProxy CA using a frontend server which is both an OpenID Connect provider and an OpenID Connect protected resource. For further information, see Section 3.22 of Milestone MJRA1.1: Existing AAI and available technologies for federated access [MJRA1.1]

CLARIN

The CLARIN projects have "homeless" users and there is a central IdP supporting these, which outside of CLARIN would thus be classified as a guest IdP. The IdP is a simple Shibboleth IdP, with user data coming from CLARIN's Drupal website; the web portal solves simple user management tasks such as e-mail-based registration (including questions regarding applicants' expertise and affiliation to CLARIN), and password recovery. Accounts need approving manually by administrators. The CLARIN IdP will always release the attributes eduPersonPrincipalName, commonName, mail, organisation, eduPersonScopedAffiliation (with the fixed value "member@clarin.eu"), eduPersonEntitlement (asserting an academic vs. public e-mail address) to CLARIN SPs; attributes released to other SPs – where the CLARIN IdP is a guest IdP – need to be agreed.

DARIAH

The DARIAH project runs its own IdP, as many users of DARIAH are "homeless," i.e. they are not members of an organisation in a national authentication infrastructure or eduGAIN. DARIAH's own SP requires the release of at least eduPersonPrincipalName; the DARIAH IdP will also release further attributes to the service via direct SAML Attribute Queries, such as email, attributes regarding acceptance of terms of use, and authorisation attributes such as privilege group memberships, that are consumed by DARIAH services for making access control decisions. All such further attributes can also be managed for users who authenticate with their institutional IdP. These typically do not have such information (community group memberships, acceptance of DARIAH AUP) or often are not willing to share personal data (email). DARIAH provides a self-service interface that allows requesting an account, for changing the password, for getting a new password, when the user has forgotten the old one, for consenting DARIAH specific AUPs, for registering additional attributes such as ORCID ID, and for applying privilege group memberships. An administration interface eases the distributed management of the data and the granting of the account and group management requests.

The DARIAH guest IdP, which is included into the DFN AAI and thus also into eduGAIN, is operated according to a documented and publically available policy for checking the identities relating to participation in the research community and for keeping the data current. DARIAH also has a sustainability strategy for all services, including the operation of the IdP.

EGI AAI Platform

The EGI AAI aims to enable users to access EGI Services/Tools using federated authentication mechanisms, for example using credentials provided by the IdP of their Home Organisation through eduGAIN. Users who do not have an account on one of the federated IdPs, should still be able to access the EGI services using social identity providers, eIDs, or other selected external identity providers. To achieve this, the EGI AAI has built-in support for SAML, OpenID Connect and OAuth2 providers and already enables user logins through Facebook, Google, LinkedIn, and ORCID. Support for user authentication through GitHub and Microsoft Live Connect is also underway. Each external IdP has a LoA assigned to it that is conveyed to the SP through the eduPersonAssurance attribute and the Authentication Context Class of the SAML authentication response. The EGI AAI currently distinguishes between three LoA levels, namely, Low, Substantial and High. For example, all social identity providers are assigned the Low LoA. Some EGI SPs have been configured to provide limited access (or not to accept at all) credentials with the Low LoA.

EUDAT – B2ACCESS

EUDAT runs a federated identity service called B2ACCESS. It is thus an actual service in its own right with helpdesk queue and support. B2ACCESS "consumes" external identities include social media identities (Facebook, Google, Microsoft Live, and ORCID), as well as those from selected communities (e.g. CLARIN), and in turn issues "harmonised" credentials which can be used to authenticate to EUDAT services. Each external IdP has (or will have) a LoA assigned to it, which in turn is communicated to the SP as an attribute, so an SP may choose not to accept credentials with too low an LoA, or from an IdP which it does not trust. As B2ACCESS has joined eduGAIN as a service provider, it needs to comply with the eduGAIN code of conduct of attribute management. This restriction also limits the use of B2ACCESS, as in most cases attributes cannot be passed to SPs outside of EUDAT - at least not without the user's consent. The schema used is mostly X.500 (under the OID arc 2.5.4) with a few additions. EUDAT will also need to make use of attributes from multiple sources and therefore researches how to manage attributes with different LoAs.

Globus

Globus (www.globus.org) started out as a comprehensive collection of middleware for grid computing, following the vision of Foster and Kesselman (now known as the Globus Toolkit). Today, it has rebranded itself as a user-friendly file transfer service, which transfers data between endpoints using [GFD.47]. While GridFTP uses X.509 certificates - or GSI proxies - as credentials, users log into the Globus data transfer portal using either username/password - the "Globus id," www.globusid.org - or institutional IDs (US only), or selected communities (LIGO, WestGrid), or Google.

ORCID

ORCID is a community-driven effort to provide researchers worldwide with persistent identifiers, irrespective of their current, (presumed temporary) affiliation, thus allowing a reliable and consistent way of linking publications and scientific work to authors throughout their career. ORCID is not meant to be used as an authentication service, but is included here as it provides a consistent global *identity* (as defined in section 1.1), namely the ORCID ID, which is used to link the research activities back to the researcher. ORCID provides both a registry for users to obtain the ORCID ID and manage a record of activities associated with it, as well as APIs supporting inter-systems communication and authentication. Once registered and logged in to the ORCID portal, a researcher can provide information about their education, studies, employment, funding, current

work, and decide whether this information will be made publicly available to everyone, to trusted parties (organisations) only, or kept fully private.

The ORCID portal providing access to the ORCID registry has recently (February 2016) been made available via the eduGAIN interfederation, so that users logging in via eduGAIN IdPs can link their existing ORCID account to their eduGAIN federated-access based account, thus obtaining federated access to the ORCID registry. This means users can access the ORCID registry as a standard eduGAIN SP.

An additional functionality is being introduced in ORCID, with testing in collaboration with AARC. The functionality comprises features that enable institutions, after the association of the two accounts, to add ORCID IDs to their user directories and IdM systems, and thus assert these attributes via their organisational IdPs. Institutions can also interact with researcher ORCID records via the member API [ORCID] to add or update activities, including verified education and employment, read trusted-parties data in the ORCID records, and support user's claims, such as "user verified affiliation" (cf. section 2.1.4).

Umbrella ID

Umbrella ID is the SAML-based authentication provider for several science collaborations in the European community, notably the photon and neutron PaNdata collaboration and the Instruct and BioStruct-X structural biology collaborations. Fourteen photon and neutron facilities use Umbrella ID to supplement their own identity management system, for instance to provide consistent identities across the facilities. In this context, Umbrella becomes the guest IdP: while its use was intended for the PaN (photon and neutron) community, it is used as a supplementary attribute provider in the facilities' user databases.

Designed to be lightweight, Umbrella ID exposes only the least amount of personal information necessary to comply with data protection regulations in all intended jurisdictions. Umbrella ID releases only three attributes:

- uid - The (globally unique) username of the Umbrella ID user involved.
- EAAHash - A globally unique persistent identifier used for matching an existing Umbrella account with an existing facility account. The EAAHash can also be used on its own as a globally unique persistent pseudonymous identifier.
- EAAKey - A key used together with the above EAAHash attribute in a challenge-response handshake to verify whether a user is registered at a specific facility.

Umbrella ID also supports linking with other SAML federations through its eduGAIN bridge. This functionality is used to log into resources that are not published to eduGAIN but do support Umbrella ID, or conversely, where the user would prefer to log in with their Umbrella ID instead of their institutional ID. A one-off account linking process at the Umbrella ID website allows a user to link their eduGAIN-registered institution ID with their Umbrella ID (through the released Persistent ID), and then use their institution ID to appropriately release their Umbrella ID.

Signing up with the Umbrella ID will require only a minimal amount of information, and verification of the account occurs by email. Manual approval of accounts by an administrator is also possible. Multi-factor authentication is not available, although certificate-based login has been explored. User-managed attribute release is not available, as the only attributes released are all required by the service providers in the Umbrella ID ecosystem. In practice, each research facility will record the user's Umbrella ID but will not rely

on it solely for access; they will always check the user's passport or other relevant ID as a prerequisite for authorisation.

Umbrella ID usage currently numbers in the low four figures, although more uptake is expected. The technologies used by the Umbrella ID service are proven to be scalable into the tens of thousands of requests per minute (with appropriate resources). Umbrella ID itself also runs in a high-availability configuration managed with geoDNS to spread load.

Zooniverse

Zooniverse is a common AAI system behind a set of "citizen science" projects [Zooniverse]. The idea is that people can sign up with a single account and contribute to science projects of their liking/interests, so their identity is usable in a large number of projects (albeit currently only projects already in Zooniverse). As such, Zooniverse defines a loose community of citizen scientists, but the AAI also supports a form of RBAC: for example, a project owner can specify "experts" who create the training data sets, "researchers" who consume the "science output" and can answer science questions from the public, and "collaborators" who can edit the experiment workflow.

References

[AARC]	https://aarc-project.eu/documents/
[Confluence]	https://www.atlassian.com/software/confluence/question
[ENISA1]	Managing Multiple Electronic Identities, ENISA https://www.enisa.europa.eu/publications/mami
[ENISA2]	Mapping Security Services to Authentication Levels, ENISA https://www.enisa.europa.eu/publications/map-auth-lev
[ENISA3]	Privacy and Security Risks when Authenticating on the Internet with European eID cards, ENISA, https://www.enisa.europa.eu/publications/eid-online-banking
[ENISA4]	Privacy Features of European eID Card Specifications, ENISA, https://www.enisa.europa.eu/publications/eid-cards-en
[ENISA5]	Security Issues in Cross-border Electronic Authentication, ENISA, https://www.enisa.europa.eu/publications/xborderauth
[FIM4R]	http://cds.cern.ch/record/1442597?ln=en
[GDPR]	http://ec.europa.eu/justice/data-protection/reform/index_en.htm
[GFD.47]	I Mandrichenko, W Allcock, T Perelmutov: GridFTP v2 protocol specification, http://www.ogf.org/documents/GFD.47.pdf
[GFD.225]	D. Groep, M Jones, J Jensen (eds): *IGTF Certificate Profile*, http://www.ogf.org/documents/GFD.225.pdf (to appear)
[IGTF]	https://www.igtf.net/
[MJones]	Mike Jones: *Identity & Access Management using Social Networking Technologies*, JISC project final report, April 2007.
[MJRA1.1]	Milestone MJRA1.1: Existing AAI and available technologies for federated access https://aarc-project.eu/wp-content/uploads/2016/01/MJRA1.1-Existing-AAI-and-available-technologies.pdf
[NIST]	*Security Requirements for Cryptographic Modules*, http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf
[ORCID]	https://members.orcid.org/api
[UKID]	Identity Cards, Home Department, UK Government (2004), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/251061/6359.pdf
[X1]	http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.257.01.0073.01.ENG
[X2]	https://ec.europa.eu/digital-agenda/en/blog/first-big-step-eidas-implementation-accomplished
[Zooniverse]	https://www.zooniverse.org/

Glossary

AAI	Authentication and Authorisation Infrastructure (*i.e.* the parts of an e-infrastructure or cyberinfrastructure that support authentication and authorisation)
AD	Active Directory (Microsoft)
AUP	Acceptable Use Policy
B2ACCESS	The AAI service of EUDAT (*q.v.*)
CA	Certification Authority, an authority issuing X.509 certificates to a PKI.
CoC	Code of Conduct – used about the eduGAIN Code of Conduct for services.
DFN	Deutsche Forschungsnetz, the German NREN
DNS	Domain Name System
EGI	European Grid Initiative; www.egi.eu
eID	"Electronic ID," essentially the same as ID (*q.v.*) but used primarily about government (or banking) IDs.
eIDAS	See reference [X1] and section 2.2.3.
EUDAT	www.eudat.eu ; see Appendix B.
FIM4R	Federated identity management for research; https://rd-alliance.org/groups/federated-identity-management.html ; see also [FIM4R]
FIPS	Federal (= United States) Information Processing Standard, see NIST.
GARR	The Italian NREN
GEANT	Networks infrastructure for research and education in Europe; see www.geant.org
GDPR	General Data Protection Regulation
ID	Identity or Identifier; see section 1.1.
IGTF	Interoperable Global Trust Federation – www.igtf.net
IdMaaS	Identity-management-as-a-service : see section 2.4.5.
IdP	Identity Provider – either an online service (endpoint) publishing identity attributes for authenticated users to authorised services, or, occasionally, by extension the organisation running such an IdP.
LoA	Level(s) of Assurance; see sections 1.1 and 2.1.1.
MJRA1	Milestone of JRA1; see https://aarc-project.eu/documents/milestones/
NIST	National (= United States) Institute for Standards and Technology, www.nist.gov
NREN	National Research and Educational Network
OID	Object IDentifier, see for example www.oid-info.com
ORCID	www.orcid.org ; see Appendix B and section 2.1.4.
PGP	Pretty Good Privacy, a peer-to-peer PKI
PKI	Public Key Infrastructure; two models exist – authorities (see IGTF, CA), or peer-to-peer (PGP)
RBAC	Role Based Access Control
RFC	Request For Comment; a list of documents from the Internet Engineering Task Force (www.ietf.org)
RP	Relying Party – a service which consumes and processes identity and authorisation attributes.
SAML	Security Assertion Markup Language ; see MJRA1.1.
SOAP	Originally Simple Object Access Protocol, now just "SOAP" as a word, not an acronym; a web services protocol.
SP	Service Provider (SAML)
VOMS	Virtual Organisation Membership Service; see MJRA1.3.
W3C	World Wide Web Consortium – www.w3c.org
X.500, X.509	A set of standards for attributes, resp. digital certificates, from the International Telecommunications Union (ITU).