“Advancing technologies and Federating communities”

Study on Authentication, Authorization and Accounting (AAA) Platforms For Scientific data/information Resources in Europe

Contents

Executive Summary

Supporting and promoting scientific research and innovation, as well as enabling access to scientific information, have always been key priorities for the European Commission and the Member States. It is widely acknowledged that Authentication and Authorisation Infrastructures (AAIs) play a crucial role in supporting research and in providing a distributed virtual environment where scientific resources can be stored, accessed and shared. More interactive, collaborative approaches to research in conjunction with the deluge of data are opening new frontiers to data processing, storing and preservation; this also poses new requirements and challenges for existing AAIs across Europe.

The **goal of this study**, prepared for the European Commission, is to evaluate the feasibility of delivering an integrated Authentication, Authorisation Infrastructure, AAI, to help the emergence of a robust platform for access and preservation of scientific information within a Scientific Data Infrastructure (SDI).

"The report also describes some target scenarios (for expected/suggested future use of SDI) that should allow the identification of requirements to future AAI/ AAA."

The **output** of the study consists of a set of recommendations for the delivery of an integrated AAI for the European SDI. The recommendations target different stakeholders, such as representatives within the European Commission for the definition of a possible directive; developers to encourage them to use specific standards to achieve inter-operability; Member States for creating the conditions for such an infrastructure at a national level; and policy makers, particularly those involved in the Data Protection Directive, to create awareness of the impact of legislation on cross-boundary access management.

This document focuses on three key messages:

1. Presenting the requirements for the AAI for SDI, as derived by a collection of use-cases identified among different communities. The use-cases call for:

a. federated access;
b. a trust infrastructure to motivate researchers to share/open their research environment to other researchers;
c. policies (and consequently proper authorisation mechanisms) to protect data ownership and intellectual property rights.

The 'Federated Identity Management for Scientific Collaborations' paper, is recommended for a more in-depth technical analysis of the eResearch requirements.

2. Analysing the results from a state-of-the-art survey of existing AAIs. Investments have been made over the last ten years to deploy AAIs to serve different purposes; examples of this are eduroam, eduGAIN, EGI, and PRACE. The overview and analysis provided focuses on the infrastructures currently used in the research and education sector, their underlying technologies and standards and the use-cases they support. This section of the study provides a high level overview of the implications of data protection

laws on the international transfer of personal information. A more extended document is available online. Because of the diversity of the requirements coming from the various communities and because of some limitations within the current technologies, it is impossible to have a one-fits-all infrastructure. However some trends can be observed:

a. All infrastructures evaluated provide Single-Sign-On for the users, although the technology used varies;
b. No single AA technology can be universally adopted, but there should be mechanisms in place to allow for integration of different technologies;
c. There is an increased interest in using identity federation, although enhancements to the current identity federations are needed to better address eResearch requirements;
d. There is an increased interest in cloud computing, which is considered as a cost-effective solution for the data deluge problem, however there are still security considerations that need to be addressed.

3. Presenting the main challenges and recommendations that the European Commission and other relevant stakeholders should address to develop an open and sustainable AAI for the SDI. The recommendations have been organised into:

a. Technical Recommendations;
b. Policy and Practice Recommendations;
c. Legal Recommendations;
d. Recommendations for Funding Agencies, EC and Member States.

The general assumption confirmed by this study is that an AAI for SDI should be built on standard technologies and that identity federations play an important role; however more research is needed to improve authorisation and accounting mechanisms.

Support for the development of a common policy and trust framework for Identity Management is needed. REFEDS (Research and Education FEDerations) the international body led by TERENA to coordinate Identity Federation processes, practices and policies and to discuss ways to manage inter-federation work, should play a pivotal role in this process. REFEDS should evolve to become the equivalent of the IGTF within the identity space. Collaboration and communication between REFEDS, the European Commission, IGTF, eIRG and ESFRI, libraries should be improved; dedicated funding to support this should be provided.

Lastly, consistent implementation and interpretation of the legal requirements in the Data Protection area is essential when building an international infrastructure.

The current version of the report will be slightly adapted before full publication to include inputs received during the final workshop scheduled on 12 July 2012 in Brussels.

1 Motivations for the Study

1.1 Background

Supporting and promoting scientific research and innovation as well as enabling access to scientific information have always been key priorities for the European Commission and the Member States. Rapid developments in Information and Communication Technologies (ICT) have made the Internet much more pervasive and have changed the way in which researchers work. Scientific research has become extremely data intensive and much more interdisciplinary, international and real-time.

A consequence of these changes is the deluge of data generated by scientific experiments in various disciplines, produced by wide scale observational data collection, as well as digitisation of content in the arts, humanities and sciences in general. For example, the Large Hadron Collider built to advance research in the area of particle physics will produce roughly 15 petabytes (15 million gigabytes) of data annually. This is just the data generated by one research activity in a single discipline. The genome research at the cutting edge of modern research requires access to a data volume of terabyte scale, assurance of data integrity, and around the clock data availability. It is also estimated that digitising the whole of the currently available paper based content and artifacts in humanities (history, literature, behavioural science) and art in the future will produce around 2-3 petabytes of information monthly.

Over the years, thanks to the funds made available by the European Commission, researchers have enjoyed a high speed network (provided by GÉANT), an infrastructure to access supercomputing resources (offered by PRACE), a federated wireless infrastructure to allow for seamless network access (eduroam) and online tools (i.e. wikis, chats etc) to create, share and consume digital information in real time.

Whilst there is no singular European coordinated data infrastructure serving multiple disciplines, there are a number of projects that offer a solution for specific user communities. Examples of these projects are EURO-VO, which offers access to astronomical data archives, OpenAIRE, which supports the implementation of Open Access publishing in Europe, and APARSEN, which is concerned with the preservation of the record of science.

Projects such as ODE, which engages with different stakeholders to work towards an interoperable data sharing and preservation infrastructure, show that libraries and data centres are committed to providing access to, organising, linking and storing research data in a trustworthy environment. This project also highlights that the potential arising from data deluge can only be unlocked by complementary network and computational facilities supported by interoperable data sharing, reuse and preservation services.

The data produced by research are very heterogeneous, as is the demand to access, store, protect and preserve them. This clearly represents both an opportunity and a challenge. Whilst technologies for 'Big Data' advance and empower users to access an unprecedented abundance of content, they also raise issues concerning authenticity, quality of data and copyright for existing e-Infrastructures.

" Libraries can be the conveners that establish a common ground among different players. Collaboration and partnering are essential in the eResearch environment. While some organizations will specialize in building tools and others in building relationships, both are required."

Rick Luce, 'No Brief Candle: Reconceiving Research Libraries for the 21st Century'(2008)

"... to collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed."

As acknowledged by the Digital Agenda for Europe 2020 and the e-Infrastructure Reflection Group (e-IRG) white papers, there is a need for harmonisation of existing e-Infrastructures. The High Level Expert Group (HLEG) goes one step further in the "Riding the Wave Report" by stating that: "*to collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed*".

To unlock all the benefits of data centric research for the knowledge society, Europe needs to build a modern Trans-European Scientific Data Infrastructure (SDI) to integrate existing Research Infrastructures, connect all scientific communities to a high performance network, and provide access to high performance computing. New types of (data centric) infrastructure require new types of access control and security infrastructure capable of answering the challenges of data persistency, authenticity, long term preservation, and privacy.

All signs point in the same direction: the underpinning infrastructures to rapidly transmit (high speed networks) and process data (high performance computing facilities) should evolve into a next generation infrastructure that offers scientists and citizens alike the opportunity and means by which to harness the potential of data. What this infrastructure should look like and the conditions necessary for its implementation are key questions which this study sets out to answer.

1.2 Objectives of the Study

The goal of this study, prepared for the European Commission, is to evaluate the feasibility of delivering an integrated AA(A)I to help the emergence of a robust platform for access and preservation of scientific information (SDI).

The goal has been broken down into two objectives:

1. A collection of **user requirements** coming from different communities concerning access and;
2. A **gap analysis of the existing AAIs** used in the realm of research and education, the use-cases they support and the associated challenges.

The **output** of the study consists of a set of recommendations for the delivery of an integrated AA(A)I for the European SDI. The recommendations target different stakeholders, such as representatives within the European Commission for the definition of a possible directive; developers to encourage them to use specific standards to achieve inter-operability; Member States for creating the conditions for such an infrastructure at a national level; and Policy Makers, particularly those involved in the Data Protection Directive, to create awareness of the impact of legislation on cross-boundary access management.

Because of the multiplicity of requirements, such as support for different user communities across different countries, support for cross-disciplinary data sharing to protect data integrity and ownership, and support for different access levels, the AA(A)I for the SDI needs to be designed to offer flexible and scalable access control mechanisms. Clearly this AA(A)I has to ensure that resources and facilities are used in the correct way and that data is accessed by users authorised to do so. It is also important to ensure that policies are implemented to deliver a trusted environment for researchers.

Figure 1 shows the organisation of the study, the role of the partners and experts in relation to the study objectives, and the final output of the study.

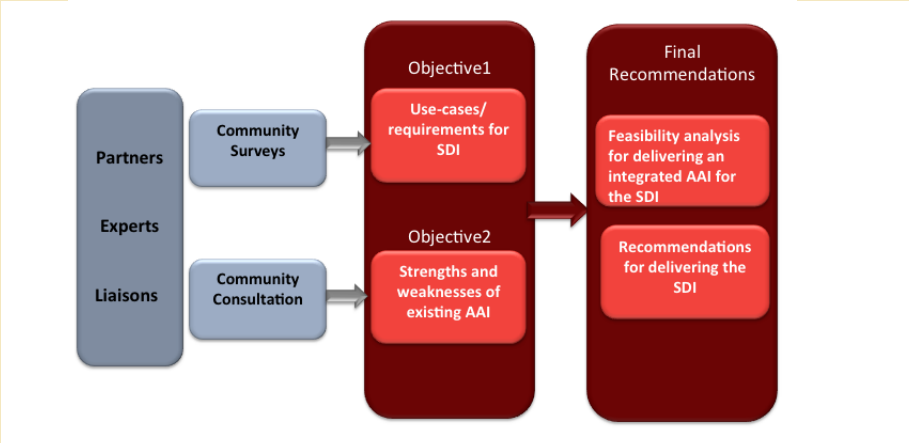


Figure 1: Organisation of the Study

1.3 Methodology

Utilising the diversity and strengths of the partnership, the study took a dual approach to exploring the challenges and opportunities associated with implementing an Authentication and Authorisation Infrastructure for the future SDI:

1. **Use-cases** have been derived from interviews with stakeholders within the e-Science, networking and library communities. The selected use cases reflect issues such as data sharing, persistent access, data curation, data management and governance, and long-term preservation. These interviews and use cases have been instrumental to assessing how existing initiatives can meet the resulting requirements and in describing future scenarios that would benefit from the SDI.

2. **Existing and emerging infrastructures in the realm of research and education have been surveyed** in order to assess how well they meet the requirements identified through the use-cases. The survey provides a complete overview of the AAI landscape in Europe as well as identifying interoperability features.

" Use-cases have been derived from interviews with stakeholders within the e-Science, networking and library communities."

Finally the study proposes recommendations for the integration of existing research and education e-Infrastructures in order to build an appropriate AAI for the SDI. It highlights issues and identifies technical, organisational, regulatory and legal obstacles to pan-European AAI platforms.

FIgure 2 depicts the approach followed by the partners in delivering the final recommendations.

Figure 2: Methodology

Community feedback has been sought on all preliminary results of the study. Draft reports have been circulated to the partners communities for review. The results presented in this final report have been validated and can therefore be considered representative of the current situation.

1.4 The partners

The study, led by TERENA, has been carried out by four partners representing the networking, library and eScience communities (see table below). The partners have been supported by external experts throughout the study. These experts provided contributions on specific topics as well as general comments on the overall study.

The strength of the study lies in the diversity of the partners involved and the variety of expertise they contribute. The four partners combined provide access to a wide cross-section of stakeholders and relevant networks as well as a unique insight into the issues, both human and technical, associated with implementing and deploying an Authentication and Authorisation Infrastructure in cutting edge environments.

" Central to this study is an understanding of the human aspects of research information access and use."

Central to this study is an understanding of the human aspects of research information access and use. Libraries have been a traditional intermediary between researchers and sources of research information. With the growth in Open Access mandates and increasing need for digital preservation services, many libraries have also been tasked with the management of institutional repositories. This has provided insight into not just how researchers access information, but also the barriers and drivers behind data deposit. This study marries the insight of the library and related research infrastructure communities with the technical expertise of groups already active in the AAI area.

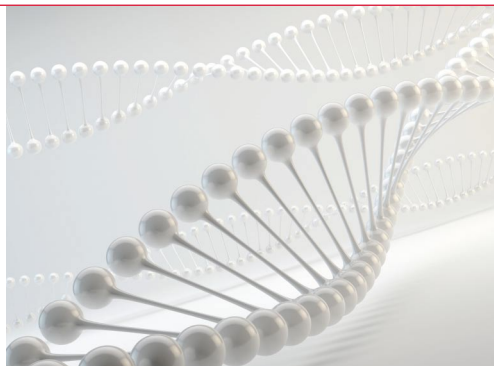
Partners	Description
TERENA Trans European Research and Education Networking Association	TERENA, the association of National Research and Education Networks (NRENs) in Europe, has approximately 40 members including international members – CERN and ESA - a number of associate members and a few industrial organisations. The mission of TERENA is to offer a forum to collaborate, innovate and share knowledge in order to foster the development of Internet technology and services to be used by the research and education community. TERENA is involved in eduGAIN, eduroam and other GEANT activities. Licia Florio has coordinated the overall study on behalf of TERENA and has provided contributions concerning the survey of AAIs and the final recommendations.
LIBER Association of European Research Libraries	LIBER is the main research libraries network in Europe. It has over 430 members from national, university and other research libraries across 45 countries. LIBER is actively working to promote the role of libraries within the European research infrastructure; in digital curation, research data sharing, and Open Access. Susan Reilly has coordinated the user-community requirements for the library sector.
UvA University of Amsterdam	The System and Network Engineering (SNE) Research group at the University of Amsterdam researches cross-domain interaction between Grid resource providers, optical and hybrid networking, resource descriptions using semantic web, and programmable networks for the Future Internet. SNE has expertise in Cloud architecture and security infrastructure research and development, generic AAA architecture and AAA framework implementation. Yuri Demchenko has provided input from the eScience perspective.
DEENK University and National Library of Debrecen	DEENK is the largest research university in Hungary and has a strong tradition of international collaboration in science and scholarly communication. It provides technical support for the Hungarian Open Repository Network. It also hosts a digital archive for PEER (Publishing and the Ecology for European Research). Tamás Varga and Gabriella Harangi have run the libraries surveys.

Table 1: Partners

External Experts	Expertise
Nicole Harris (JISC Advance)	Nicole Harris has extensive experience within the education sector as an advisor and project manager with a focus on access and identity management. Nicole has coordinated the user-communities requirements concerning the networking community and has provided her expertise in finalising the recommendations.
Diego Lopez (Telefonica I+D)	Diego Lopez has been involved in a number of projects and initiatives at RedIRIS and TERENA while working at RedIRIS; he chaired the TERENA Task Force on Middleware (TF-EMC2). Currently working at Telefonica I+D he focuses on advanced network infrastructures and security. He contributed to the 'Riding the Wave' report as one of the experts in federated access technologies. Diego has reviewed the survey on AAIs and provided feedback on the final recommendations.
Klaas Wierenga (Cisco Systems)	Klaas Wierenga is considered the 'creator' of eduroam, the federated infrastructure for network access. He has been involved in several projects both as a SURFnet employee and in his function within Cisco Systems. Klaas is currently chairing the TERENA Task Force on Mobility and Network Middleware. At Cisco he focuses on security, (federated) identity and mobility. Klaas contributed the eduroam and Project Moonshot sections of the AAI survey and informed the final recommendations for the report.
Torbjörn Wiberg (Univ. of Umeå)	Torbjörn Wiberg has been involved in deploying authentication and authorisation infrastructures at campus level. Torbjörn has been also actively involved in the eInfrastructure Reflection Group for several years.
Andrew Cormack (JANET)	Andrew Cormack works as Chief Regulatory Adviser, dealing with regulatory and policy issues of running and developing the network and its services. Andrew advised the study team on Data Protection topics.

Table 2: Experts

2 Community Requirements



2.1 Introduction

If the purpose of a **Scientific Data Infrastructure** (SDI) is to enable researchers to create, store and share the data resulting from their experiments, and to find, access and process the data they need, then the practices and concerns of researchers must be central to the definition of requirements for this study. An SDI and any plans or directive for its implementation must accommodate current practices, address current and future needs, and take into account the concerns researchers and information providers have in relation to data sharing and the use of an SDI.

There are several efforts already underway to investigate how existing Authentication and Authorisation Infrastructures (AAIs) can be adapted or extended to better meet the requirements of diverse research communities: an excellent example of this is the 'Federated Identity Management for Scientific Collaborations' paper, coordinated by CERN with input from a range of research organisations. The New Global Data Generation Manifesto, signed by several member of the European Strategy Forum for Research Infrastructures (ESFRI), highlights the urgency of the need for authentication infrastructure to be is in place at the institutional and national levels and for the harmonisation of related policies.

" An SDI and any plans or directive for its implementation must accommodate current practices, address current and future needs, and take into account the concerns researchers and information providers have in relation to data sharing and the use of an SDI."

2.2 The Data Sharing Community

The data sharing community is heterogeneous in nature. Their requirements vary according to discipline, the nature of the data to be shared, its scope for reuse, methods for collaboration, and culture. It is not within the scope of this study to provide in depth analysis of the type of requirements which could be drawn from individual disciplines, as some disciplines and data sets have very complex and unique requirements. However, some more generic issues can be drawn from a brief analysis of the high level requirements of broad disciplinary areas. The following are the scientific areas defined by the ESFRI and an outline of their community requirements for an AAI:

- In Biological and medical sciences, a discipline that generates huge volumes of data, issues of data sensitivity are common and any data sharing must adhere to data privacy laws and policy. In the social sciences and humanities, whilst the data may be less sensitive, they may still be subject to license and issues around the 'long tail' of managing smaller data sets;
- Environment and earth sciences have a strong tradition of data sharing, generate high volumes of data, and are more advanced in terms of the technology used to exploit and interact with data than other disciplines;
- Material science, analytical and low energy physics is characterised by short projects and experiments, leading to a highly dynamic user community. This community would benefit from a collaborative infrastructure that would allow for both local participation and remote access. The community expresses interest in using federated identity management to reduce administrative overhead and in tools to manage ad-hoc collaborative user groups via virtual organisations or federations;

- Lastly it is worth mentioning the ENVRI project and LifeWatch project as examples where data sharing amongst different disciplines is a key enabler for system-wide science. The main challenge for these groups relates to data capture from distributed sensors, metadata standardisation, management of high volume data, workflow execution and data visualisation.

2.3 The Nature of Data

The Data Publication Pyramid (fig.1) illustrates a growing problem, which an SDI and AAI could help address. The pyramid visualises the ways in which research data can be made available. The base of the pyramid represents data stored locally in its raw form on hard drives and disks; these data are typically the result of scientific experiments or analysis. There are several reasons why these research data are not shared, varying from intellectual property protection concerns, to ethical, technical or cultural reasons. An AAI is one of the mechanisms that can help improve data sharing practices amongst researchers through providing technical support for addressing data access control and policy related issues. An AAI could make sharing data more straightforward technically and safer for research and also help ensure the integrity and security of the data. It could also be argued that by making data sharing more simple and potentially more common place, an AAI will facilitate a cultural shift in terms of data sharing and collaboration.

The second layer of the pyramid visualises data which is already stored in repositories. This data is available for use and reuse. Here an AAI can facilitate collaboration and provide authorised access to the wider scientific community. Although a certain amount of this data may be open, some of it may require authorisation for ethical, regulatory and security reasons. It is also the case that open research data may be linked to by licensed or copyrighted publications and vice versa.

The third layer of data shows publications as supplemental files to articles. Again, these articles and files are located on a publisher platform and can be open access or licensed. The top layer is the traditional view of an article or publication with the data embedded within.

For a truly collaborative data infrastructure to take effect researchers need to be able to work with, and collaborate using, data from all of these layers. They need to be able to share their data easily and securely and they also need access to licensed and open access content, from data sets to the finished article.

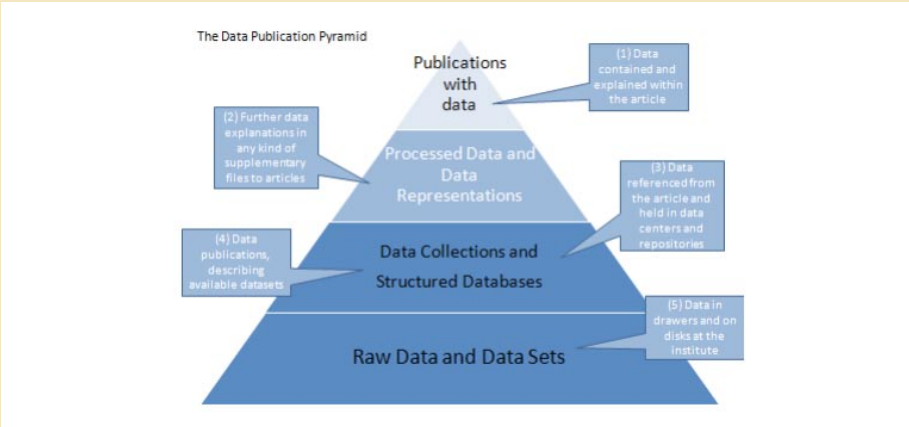


Figure 1: The Data Publication Pyramid (source ODE Report in Integration of Data and Publications)

The entire data lifecycle (fig.2), which covers different stages spanning from data collection and filtering, data analysis, data publishing, data archiving, and also includes additional stages such as archived data discovery, re-purposing and re-using, needs to be taken into account. The future SDI should support all the stages in the data lifecycle and allow for multipurpose data collection, use and advanced data processing. Facilitation of the storage of initial data sets and all intermediate results will allow for future data use, in particular data re-purposing and secondary research, as the technology and scientific methods develops. These functions may become part of a new role for libraries as curators and managers of more complex scientific data, data linkage and new networks of resource sharing.

The future AAI must work seamlessly across the whole data lifecycle and address focused issues, such as general data security, access control, data usage policy and related Intellectual Property Rights (IPR), data linking, data filtering, secondary data mining, and data and research information archiving requirements.

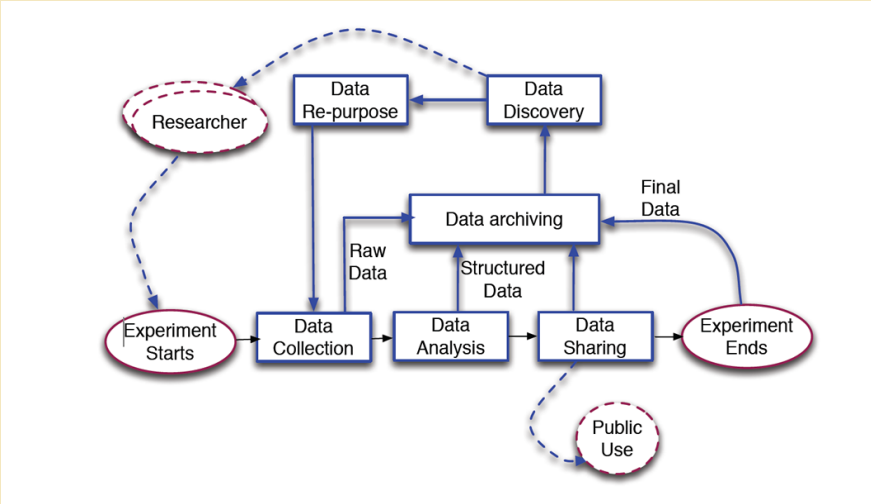


Figure 2: Data Life Cycle Model (source Data Lifecycle Models and Concepts)

2.4 What do we know about Researcher Requirements?

Current practice in relation to how researchers find, access and process research information and data has been drawn from a survey sent out to research librarians from the 430 libraries within the LIBER network and other communities. The libraries received two surveys. One survey was aimed at the librarians themselves and contained questions relating to how their resources were authenticated and the behaviour of their users. The other survey was sent to the libraries' research communities and explored practices and preferences relating to authentication, as well as attitudes towards the use of information resources and data sharing.

Roughly 100 librarians responded to the survey and the response represented a fairly even geographical spread. The researcher survey had 600 responses. Neither survey was designed to be statistically representative of nationality or discipline, but it does provide an insight into relevant practices and attitudes.

The survey reveals the following:

Researchers primarily use their institutional credentials for authentication (fig. 3), although a not insignificant number (19%) use their social network account credentials to access scientific information.

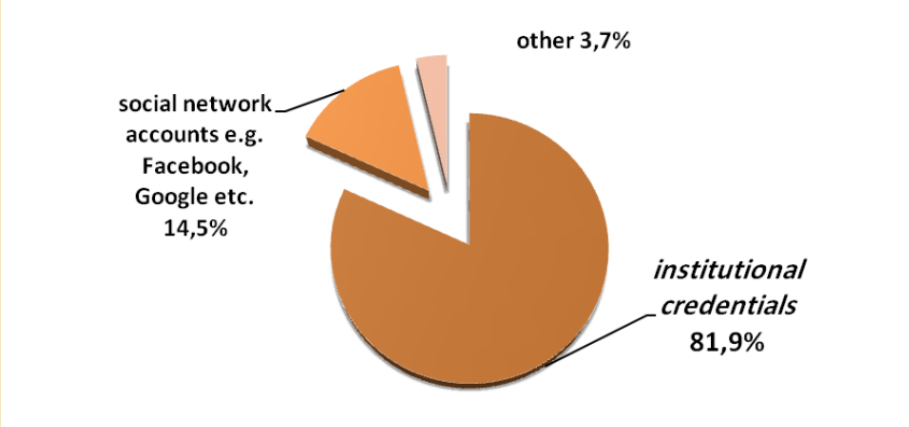


Figure 3: Credentials used by researchers

Nearly half of researchers use more than one credential, but a large majority would prefer to access all resources using their institutional credentials.

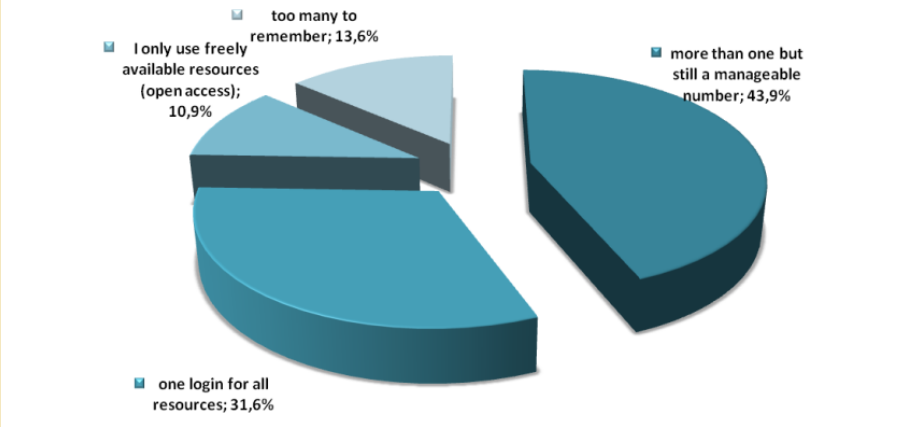


Figure 4: Number of credentials used by researchers

IP-based authentication (fig. 5) is still the most widely used method of providing researchers with access to information resource subscribed to by institutions.

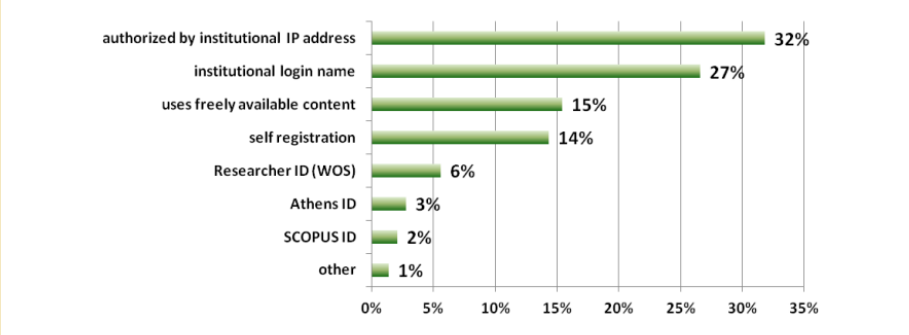


Figure 5: Access to institutional subscriptions

The number of researchers depositing or sharing their data in repositories is growing (fig. 6) but there are a large percentage of researchers that are still not depositing data. This is down to issues such as trust, IPR and also the fact that researcher cannot always deposit their material directly - they must go through an intermediary e.g. a librarian.

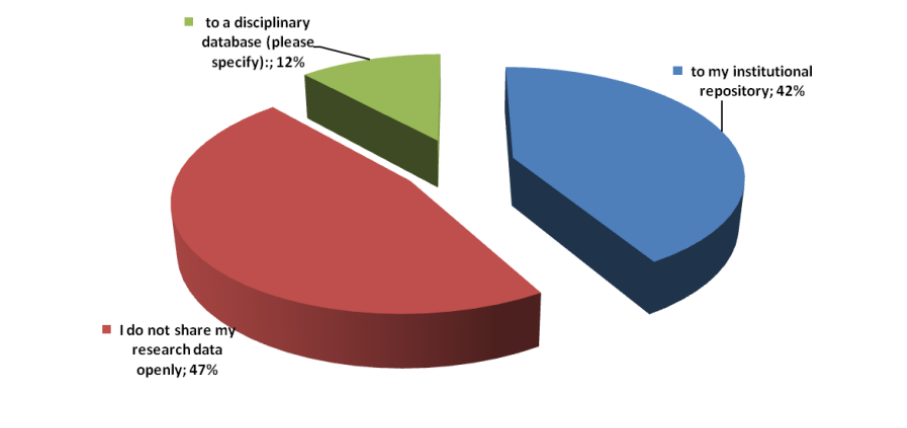


Figure 6: Depositing of data

2.5 Accessing Scientific Data and Information

Currently, researchers are largely reliant on institutional credentials for authentication purposes but it is possible that, as a new generation of researchers enter the fray, the percentage of researchers using social media to share and source information, and for collaboration, will increase.

As the survey indicates, the use of IP addresses to identify institutional users is still the common practise for accessing online material to which institutions subscribe. This method, characterised by its simplicity of implementation, allows users to seamlessly access resources without needing to go through an explicit log-in process. However this method, has serious shortfalls as it neither identifies the user, nor provides permissions but instead ascertains that the device used to access is in a certain IP range at the time of access. Beside the security considerations (i.e. IP addresses can be easily forged), it limits users to access a resource on campus, which is clearly a limitation in today's mobile world.

Despite its large adoption, this approach does not provide any features that are desirable in any AA(A)I for the SDI; it is in fact envisaged that systems like Shibboleth will completely replace IP-based access as a means of providing access to information behind paywalls.

The requirements to move towards open access, which would mean that more publications become publically available is emerging in various communities. This requirements will need to be supported by open access technologies that will require new functionalities, beyond the current institutional subscription or individual paid access to licensed content, to be implemented within the AAI.

Persistent identifiers will play a more important role for data and for researchers as authors and users of data. Persistent identifiers will enable an 'online identity trail' and will improve accounting and statistical analysis of scientific information, publication usage and their inter-relation; this is important both in terms of ensuring that researchers as creator get recognition for their work and in terms of the preservation of the record of science.

In order to discover the relevant current and future needs and concerns of the researcher in data sharing, use cases have been drawn from consultation with stakeholders from within the library, e-science, and research infrastructures communities. The following are a selection of the use cases collected through the study. They are a snapshot of the wide array of uses and define some of the requirements for the AAI for SDI in relation to supporting the scientific community in the use and reuse of research data.

Use-case 1
Creating Data

Researcher identification: the ORCID initiative

A fundamental question for researchers, publishers and funding bodies is how to trace a researcher's publications and other research contribution back to the correct person. Researchers frequently move between research groups, change institution affiliations, and even their name. Furthermore, the conventional way to identify a researcher is by last name and initials of the first name, leading to ambiguities since natural names are by no means unique.

The Open Researcher & Contributor ID Initiative (ORCID) is an international, cross-community and interdisciplinary initiative dedicated to solving the name ambiguity problem in scholarly communication. ORCID will work to support the creation of a permanent, clear and unambiguous record of scholarly communication by enabling reliable attribution of authors and contributors through unique identifiers.

The process of 'asserting identities' will involve, in the simplest scenario, researchers registering themselves to create a profile from scratch, then adding publications to their profile. These are claims are self-asserted, both biographical and bibliographical. A small number of universities and other research institutions will bulk-create ORCID profiles, and subsequently those researchers may claim this record from ORCID. After claiming the profile, the researcher will then have control of the profile data.

Tracking provenance is a priority for phase two of ORCID. The aim is to enable consumers of profile data to see where each piece of information came from, and decide based on the provenance whether or not to trust that claim.

Requirements:

- Tracking of provenance, authenticity, integrity of the material;
- Integration of researcher ID with institutional credentials;
- Self registration;
- Securely linking researcher and data identifiers for tracking provenance.

Use-case 2
Analysing and Processing Data

Experimenting with data: the Scientific Information Service at CERN

CERN has a staff of some 2400 people, 3/4 of these are technical people. However, the largest supported group of the Scientific Information Service is the CERN 'users': scientists (high energy physicists) from all over the world who come to Geneva to use the CERN experimental facilities (particle accelerators and detectors).

Currently the CERN Information Service has 11000 registered users (including CERN fellows). CERN currently runs the Large Hadron Collider (LHC) experiment which produces huge amounts of data which are processed by the international worldwide scientific community; this requires a robust and integrated infrastructure to manage and access data, protect data integrity and support the whole scientific data lifecycle.

The LHC experiments and community are supported by the Worldwide LCH Grid (WLCG). User cooperation and access is organised in the form of Virtual Organisations (VO) which manage user roles and issue user certificates in the form of X.509 Attributes Certificates.

The users of the Information Services at CERN come from over 600 different universities worldwide. These users would benefit from being able to access both their institutional and CERNs information services seamlessly. Information access could be made more cost effective, as currently the community often pays for access to resources twice due to complexities of resource access whilst researchers are in different locations. In this case both the home institution and CERN are paying for the same researcher to access the same resources.

A priority for the CERN Information Service is that the right to edit data is only given to authorised individuals and that any changes to data can be traced back to these individuals.

Requirements:

- Track user or organisational access to CERN resources;
- Strong authentication level, typically based on face-to-face verification;
- Entitlement management to reduce situations where resources are being paid for twice;
- Attributes for the users that participate in a specific research project should be managed by the specific research groups (VOs).

" When you actually get to the issue of offering the possibility of editing data, then it is of course highly important that this access is only given to authorised people and that all changes can be traced back."

Jens Vigen
Head Librarian, CERN

Use-case 3
Sharing Data

Working with communities of practice: DARIAH

The Digital Research Infrastructure for the Arts and Humanities (DARIAH) aims to enhance and support digitally-enabled research across the humanities and arts by developing, maintaining and operating an infrastructure in support of ICT-based research practices. DARIAH will work with communities of practice to bring together individual state-of-the-art digital humanities and arts activities across Europe.

DARIAH will operate through its European-wide network of Virtual Competency Centres (VCCs). Each VCC is centred on a specific area of expertise. VCCs are interdisciplinary, multi-institutional and international. An AAI service, along with a Persistent Identifier (PID) resolver service, are core technical services that DARIAH will guarantee as part of its digital research infrastructure.

For researchers, having easy yet secure access to all the resources, data, tools and services they need to undertake research will be a significant advantage. In today's time-poor society being able to login easily using one set of credentials will have significant time savings, rather than having to negotiate through a multitude of different authentication systems.

A European-wide single sign-on service will encourage researchers to share their work within a secure and trusted environment. For this, authorisation granularity is essential. With such a service, authorisation could be granted to a secure personal environment for an individual's own research, to secure spaces for sharing with a closed research group, or more widely to large research communities. As more, often national, AAI infrastructures are developed the technical challenges of making them work together in a seamless and user-friendly way will need to be addressed.

The sharing of research data across borders presents legal issues. Even if the technical solutions are in place, resource licenses are typically negotiated at a national level and will not allow access to resources to be authorised on a pan-European level. Conversely, being able to ensure a secure environment for Europe's higher education community may encourage resource owners to consider licensing resources at a pan-European level.

" A European-wide single sign-on service may encourage researchers to share their research, within a secure and trusted environment. For this, authorisation granularity would be essential. For example, authorisation could be granted to a secure personal environment for their own research, to secure spaces for sharing with a closed research group or for a wider research community."

Sally Chambers
Secretary General, DARIAH

Requirements:

- Authorisation granularity to groups of users from multiple organisations at a variety of levels;
- Delegated rights to authorise users to become members of specific groups;
- Increase number of participating institutions in current access management solutions.

The Life Sciences

Biological and medical sciences (also defined as life sciences) have a general focus on health, drug development, new species identification, and new instrument development. Life Sciences generate massive amount of data and pose new demands for computing power, storage capacity, and network performance for distributed processes, data sharing and collaboration. The amount of data that the life science community generates require powerful data centres with high performance network for data transfer.

There are already existing dataset generated by previous research in this field; one of the main challenges is to interlink them and enable authorised people to access them remotely. Connecting existing data to each other would require new fine grained access control policy and a consistent enforcement system.

Furthermore biomedical data (healthcare, clinical case data) are privacy sensitive data and must be handled according to the specific provisions in the European Policy on Personal Data processing for such information.

Requirements:

- A trusted environment for data storage and processing;
- User friendly data encryption;
- Fine grained access control policy;
- Possibility to filter data based on the applied policy;
- Policy binding to data in long-term storage to protect privacy;
- Tracking of data usage.

Use-case 4
Preserving Data

Enhancing Publications: the SURFshare Programme[1]

An Enhanced Publication is a new type of publication. It links typically text-based publications with additional material such as research data, models, algorithms, illustrative images, metadata sets or post-publication data like comments or rankings. The option of changing post-publication data allows an Enhanced Publication to develop over the course of time.

There are several baseline criteria for an Enhanced Publication that are of interest to an AAI, such as recording authorship of the publication and its components.

Enhanced Publications (EP) is a core activity in the SURFshare programme run by the SURF Foundation. Projects range across disciplines, from the humanities to the hard sciences. The technical infrastructure is similar across the different disciplines, facilitating easy exchange of information across systems. Different disciplines have different habits and needs; to serve these wide-ranging needs the SURFshare programme uses customised tools that support individual workflows.

[1] Adapted from Ten Tales of Drivers & Barriers in Data Sharing

" The advancement of data sharing remains a big challenge. Researchers hesitate to publish data. This is a barrier for both national and international initiatives. This raises some hard questions such as what licences should be in place? One proposition could be open access where possible, closed when needed."

Wilma Mossink
Project Manager, SURF

Repository infrastructures are being upgraded to support the creation, storage, visualisation and exchange of Enhanced Publications. A common data model is used in the development of customised tools required in the various EP projects. Eventually all Enhanced Publications will be aggregated in Narcis, the open access portal for scientific output in the Netherlands.

Another focus of the SURFshare programme is permanent access to research data. SURF started with Enhanced Publications, but quickly realised that this style of publication could not happen without proper data preservation and data access models. Licensing and related aspects play an important role in data access; furthermore there are several baseline criteria for an Enhanced Publication that are of interest to an AAI, such as recording authorship of the publication and its components.

Requirements:

- Facilitate both open access and closed access;
- Tracking of changes to data post-publication;
- Easy exchange across information systems.

Use-case 5
Analysing Data

Facilitating new research environments: Goettingen State University Library

Goettingen State University Library is a central unit on the university campus. The university provides programmes across almost all academic disciplines and incorporates the Academia of Science. The University Library is also the State Library for Lower Saxony and houses the national collection of 18th century material, and the biggest scientific library within Germany.

" A central process for authorisation and authentication needs to be agreed across centres and legal issues need to be harmonised.Researchers will not use these new research environments if they have to agree terms of use with each individual party."

Heinke Neuroth
Head of Research and Development
Goettigen State & University Library

The subject collection principle is employed in Germany and the Library has 17 special subject collections; the library collects relevant material from all over the world for these subject collections. The nature of the collections in the library attracts an international user community of researchers and scientists.

Because the collections are of international significance, both local and international researchers need access. Researchers are increasingly working collaboratively and there is a need for virtual research environments to support these groups sharing software, disk space and content. There is a need for single access to these resources and an AAI to facilitate this approach. Goettingen is one of the main supporters of the Research Infrastructures Manifesto led by the CLARIN project, which champions the need for such an AAI.

Textgrid is an example of a Virtual Research Environment used by researchers at Goettingen that allows for single sign on, but only for resources deployed by the project. This means that it is not possible to exchange research objects or have shared storage outside of the immediate infrastructure of the project.

The supercomputing centres in Germany are also considering a new approach to single sign on. Currently every researcher must agree terms of use with individual centres. There are terms of use to be adhered to at every level, from European, governmental, state to institutional, which are not easy to harmonise. These terms of use take data protection into account, as well as access to licensed content, creating a confusing picture of access rights and permissions across the centres.

Requirements:

- Access for both local and international researchers in a virtual environment;
- SSO that is not specific to single projects or platforms;
- Reduce requirements for researchers to sign-up to multiple agreements.

Use-case 6
Accessing Data

Seamless access to research information resources and repositories: University of Edinburgh Library

The University of Edinburgh Library is part of the Information Services division at the University of Edinburgh. It has a data repository service called Edinburgh DataShare. University of Edinburgh researchers access library resources in a common way, using their institutional login.

The library uses both EZproxy and Shibboleth for authentication and authorisation. There may be more than one set of credentials behind this but, for the end user, the authentication experience is the same. Users login to resources with their institutional account both on and off campus.

An example of this joined up approach to authentication is the user experience of accessing Google Scholar. If a user finds an article through Google Scholar they can use the institutional link resolver and log in using their institutional credentials. This is not always possible when accessing content directly through a publisher's website where federated access has not been deployed.

The benefits of an AAI for University of Edinburgh researchers is that it would enable easier collaboration across institutions. The value exists in the provision of access rights to the same resources and in the ability to share information. Sharing disk space and software has been achieved with relative ease but an AAI for researchers also requires a new approach to the licensing of e-resources.

An AAI could also help reduce the administrative work for authorisation of submissions to the institutional repository by allowing co-authors from other institutions to access the repository to update works.

Copyright and IPR are a major barrier to achieving the full potential of an AAI. Licensing is costly to negotiate with publishers and the solutions negotiated must be sustainable. The view of the librarians is that open access solutions are key for the success and

" That researchers use the institutional login is quite important. Users know they have the same login for their PC, VLE and institutional repository and research content. The institution login is recognisable and trusted."

Morag Watson
Librarian, University of Edinburgh

sustainability of an AAI in the long term. This necessitates investment, incentivisation and a culture change amongst researchers and institutions.

Requirements:

- Ability to collaborate across institutions;
- Support for author identifiers;
- Support for open access resources.

2.6 Summary

The review of the various user community needs shows some requirement overlap, such as the need for an infrastructure that eases collaboration whilst at the same time preserving ownership of data and authenticity. Mechanisms are needed to safely allow researchers to link their scientific results with initial data (sets) and with intermediate data to allow for future data re-use.

Research generated in such a manner creates complex intellectual and usage rights, meaning that sophisticated tools to enforce policies on how data can be accessed and processed are needed. In a highly distributed and dynamic environment, researchers (and communities) want to be sure that data held remotely are not compromised, not altered and remain under the user control.

As mentioned above in this chapter, the institutional survey showed that most of the users would like to use their institutional credentials to access services; this trends reflects the increased penetration of systems like Shibboleth, which enables federated access. Challenges and opportunities to use federated access technologies to support access to different applications are presented in chapter three of this report.

The changing approach to mass digitisation and open availability of publicly funded content and resources also implies that libraries, as intermediaries, will need to evolve as they move away from managing subscriptions and move towards enabling open scholarship and curation of data. An AAI for the SDI should take into account all the requirements and mediate between open access and the need to protect some content for ethical or privacy reasons.

In summary the AAI for the SDI should:

- Empower researchers to utilise the shared facilities of large data-centres for trusted data processing, with guaranteed data and information security;
- Motivate researchers to share/open their research environment to other researchers by providing tools for instantiation of customised pre-configured infrastructures to allow other researchers to work with multiple data sets;
- Protect data policies, ownership rights, and data linkage when providing data archiving.

3 Survey of the AAIs

3.1 Introduction

The development and deployment of **Authentication and Authorisation Infrastructure** (AAI) has taken place in different research and education environments as well as in the private and public sector. National Research and Education Networks (NRENs) have been developing and operating AAIs for over 10 years; these AAIs provide services to a great number of users within the academic and research community. The Grid community has developed their own AAI, cloud infrastructures are becoming increasingly popular and so is the need for a reliable and trustworthy cloud AAI. Lastly governments are trying to deploy an AAI to support business and citizen transactions.

An **AAI** is an **infrastructure** to verify a user's identity (**authentication**) and to verify that a user has the rights to access the service the user has requested (**authorisation**); often these infrastructures offer **accounting** mechanisms to determine how much resources users consume, to collect statistics data, to record authentication failure and other diagnostics.

The overview and analysis provided in this section focuses on the infrastructures currently used in the research and education sector, their underlying technologies and standards and the use-cases they support. Some emerging technologies and infrastructures will also be mentioned due to the impact they are expected to have on the existing AAIs. This report does not address accounting, since in the assessed AAIs only limited monitoring and statistics facilities are available.

3.2 Identity Federations

The need to access resources in different administrative domains in combination with the evolution of web technologies, collaborative and international research, and the increasing number of systems requiring authentication has imposed new requirements on access management technologies. Provisioning user accounts for each application users wish to access does not not scale well in a highly distributed and collaborative environment that crosses multiple administrative domains and national boundaries.

" Currently Identity Federations are nationally focused and their main use-case is to provide support for Single Sign On for applications that run in a web browser."

The current best practice in education and research to meet this need is based on what is known as **Federated Access**, or **Federated Access Management**, or **Federated Identity Management** or simply **Identity Federation (IDF)**.

An Identity Federation is an infrastructure where:

1. **Authentication** is controlled by the user's **Identity Provider**, also referred to as IdP (typically the institution the user is affiliated with) that verifies the user's identity and issues access credentials (i.e. username and passwords, X.509 personal certificates etc.)

2. **Authorisation** is controlled by the resource provider, also referred to as **Service Provider** (SP) or **Relying Party** (RP) that relies on the authentication done by the IdP and the information (attributes) received about that user from the IdP and possibly from other attribute providers within the Federation.

3. **Policy** or legal agreements are in place among the entities participating in the federation to achieve a trust relationship between the parties.

IDFs enable users to access applications and resources operated in different domains with the same set of credentials issued by the users' IdP; **in other words IDFs enable Single Sign On: login once to access multiple services**. This model allows institutions to offer a richer service portfolio, reducing the need for bilateral agreements between each institution and each service; typically the federation operator handles the agreements for all parties participating in the federation and provides the technical support to enable the communication among all parties.

As reported in the TERENA Compendium and by the REFEDS group, the number of federated access infrastructures in the research and education community has been growing constantly since 2005. To date the majority of the NRENs in Europe offer (directly or via a third party) federated access for their users. However the level of deployment, the participation of institutions and the amount of services available via different federations vary significantly from country to country; for instance not all research institutions, libraries and humanities centres are connected to national federations. IDFs particularly cater for users affiliated with an institution; users without an affiliation (for example because their home organisation has not joined an IDF) or users affiliated with multiple institutions or '**nomadic users**' i.e. persons who move from one institution to another, cannot be easily supported by IDF at this point in time. The 'nomadic users' pose an interesting challenge to the Identity Federations, particularly when they tend to be identified with researchers; for instance access to the researchers' publications or researchers' data may become unavailable to the owners when they move to another institution.

This problem could be solved with the introduction of a **persistent identifier** that would follow researchers when they move among institutions. Research in this area is being carried out by among others the ORCID (Open Researcher and Contributor ID) community; at the moment however there is no universally deployed solution to this problem.

Underlying Technology

The Security Assertion Markup Language, SAML2.0, is the open standard used to build IDF systems. SAML protocol supports the secure exchange of authentication and authorisation data between identity providers and service providers or relying parties. In the (higher) research and education sector the first SAML-based IDFs were introduced in 2005 by the NRENs, which have been driving the development of IDFs ever since. Well known IDF products used in the higher education sector (and beyond) are Shibboleth, the open-source software developed by Internet2, SimpleSAMLphp, the open-source community driven product developed by UNINETT and the commercial Active Directory Federation Service (ADFS) (Microsoft).

A challenge to the current IDFs comes from the rapid development of user-centric technologies widely adopted by Web 2.0 applications such as the social networks (i.e. Facebook, Google etc). Web users are becoming more and more accustomed to the access control management used in social networks; increasingly these credentials are federated: instead of creating new user accounts, users can - in principle - use their existing 'social credentials' to sign into a range of commercial third party services. The penetration of social networking has increased the demand to use social network asserted credentials for

libraries and/or university resources. In this scenario rather than using their institutional credentials users would log in using their preferred social network credential. This model has some implications:

1. The identity vetting and the authority of the information associated with these identities are still a concern; practically those identities are self-asserted and therefore the level of trust associated with them is low. However they could be used to provide access to services where authentication is not very important, but personalisation of the service is key.
2. These 'social identities' do not carry additional verified information regarding the role of the user (i.e. student, researcher, etc); this information is particularly relevant in the research and education context, to access services to apply for grant, handle users' exams and so on. This role can only be provided by the institution(s) the user is affiliated with. Therefore a mechanism to link the user social identity with an institutional identity should be in place; this approach implies a further separation between authentication of the users and effective management of authorisation rights (attributes) and would require some security mechanisms to prevent inaccurate linking.

Trust Model

The trust model in IDFs has different aspects:

1. **The Relying Party must trust the Identity Provider** that has authenticated the users as agreed and that the users information (attributes) are up-to-date.
2. **The Identity Provider must trust the Relying Party** that processes and protects any personal data received from the Identity Provider in a way which conforms to data protection laws.
3. **The users must trust their Identity Provider and Service Provider** to preserve personal information.

Clearly the user's IdP plays a very important role in vetting the user's identity (typically when the user enrolls), in updating the user's information (for instance if the user enrolls for more course, graduates etc) and in deprovisioning the account when the user departs. This model however does not address the needs of very dynamic and cross-boundary research, where researchers from several institutions working on the same discipline and participating in a research collaboration would like to maintain additional user attributes (for instance to indicate their collaboration-specific roles, group memberships and authorisations) by themselves. It is difficult to maintain these attributes in the collaborating Identity Provider, because each collaboration group maintains specific information about the participating users that are unknown to the user's Identity Provider **and collaborations often have users from several Identity Providers**.

A complementary approach is that the research collaboration, also referred to as **Virtual Organisation**, maintains the additional user attributes and has related policies, processes and tools for assigning the proper attributes to the members of the collaboration. Technically, this extends the bilateral Identity Provider - Relying Party relationship to a triangle, where the Resource Provider uses an Identity Provider to authenticate the user, and subsequently fetches his/her additional collaboration-specific attributes from a server maintained by the collaborating community, which in effect acts as Attribute Provider (AP). **Although it is generally agreed that the model above offers a solution to the problem, more research and development is needed to provide easy to use and multi-community groups managements solutions**.

Lastly there is not yet a standardised way among different federations to express that an institution has performed additional verifications on the users' identity (Level of

Assurance). The REFEDS group is tackling this problem; progresses in this area will be driving by the increasing requirement for security of the services.

In summary federated access brings a number of benefits, both for the users (reduced number of credentials, possibility of SSO), for the services (reduced bilateral agreements between the service and each institution) and for the institutions (more services can become available to the users without institutions having to operate them directly).

The table below summarises the strength and the challenges related to Identity Federations.

Federated Access Strengths	Federated Access Challenges
Enables users to access a wide range of (Web) resources, using the same credentials, known as a Single Sign On (SSO).	Moving beyond web based resources and addressing the specific needs of researcher groups (VOs) in terms of attribute management and delegation (particularly in the case in which attributes are provided by third parties attribute authorities).
Particularly useful for service providers, to relieve them from the 'burden' related to users' administration.	The deployment and the number of applications available in a federation vary from country to country; in fact not all countries offer a federation.
Good security: information about users exchanged between IdPs and RPs take place through secure channels. Furthermore the IdPs guarantee that user's personal data are protected.	Stepping up to allow stronger authentication verifications where necessary. Ensuring the appropriate Level of Assurance for provided credentials as defined by applications or resource providers.
Based on standard technologies to achieve cross-platform and cross-domain interoperability.	Working with emerging, possibly competing, standard approaches.

Table 1: Identity Federation overview

Federated Access Management and the Law

Federated Access Management involves a complex exchange of personal information between people and organisations who may not have any direct relationship, often taking place across international borders (or even involving entities whose geographical location is unclear), and often including types of identifiers that were not contemplated in 1995. It is therefore inevitable that it will present challenges of interpretation and application of data protection laws that have their origin in the European Directive (95/46/EC) of that year.

Having educational organisations act as identity and service providers for their members, as will be required for e-research, raises new legal issues because, unlike other sectors, the individual's relationship to the organisation is not just that of customer of an access management service. Any legal framework for access management in research and education must take into account the existing relationships and contracts associated with employment, education and the provision of services.

Unfortunately both justifications for processing personal data and relationships between organisations exchanging data have been implemented very differently in different Member States. **This level of divergence makes it hard to find a legal framework that will work in all Member States**; without such a common framework it seems inevitable that the establishment of international federated access management will be hindered by the different laws and expectations in different countries. **Consistent implementation and interpretation of the legal requirements across countries is essential.**

International Transfers

Many applications of federated access management in education and research involve parties outside the European Economic Area (EEA). Educational resources or teaching may be provided by publishers or universities in other countries, while research collaborations often include both researchers and instruments in other continents. Since these overseas participants are likely to play the same roles in education and research as their European peers it is highly desirable to include them within a single legal framework and agreement, rather than have to maintain two (or more) different legal arrangements among partners who are otherwise treated alike.

Unfortunately the options provided by data protection law to achieve this are very limited. Under the current Directive the only justification that can be used for transfers of personal data both within the EEA and overseas is the consent of the individual. However it seems questionable whether consent is appropriate for choices that an individual may be compelled to make to continue their employment or study. Since many current research partners are in the USA, the Safe Harbor Agreement might be an option for these, but it cannot be used by US universities as they are not covered by the relevant regulators.

Article 26(2) of the Directive permits transfers of personal data outside the EEA where the data controller "adduces adequate safeguards with respect to the protection of the privacy and fundamental rights and freedoms of individuals and as regards the exercise of the corresponding rights". Since the 'legitimate interests' justification that is being proposed for use within the EEA already requires both identify provider and relying party to ensure protection of the individuals' fundamental rights and freedoms, it appears that the required safeguards may already be in place. Although 'legitimate interests' is not currently listed in Article 26(1) as permitting exports, it seems that it might, under Article 26(2), offer a single legal framework that could cover transfers of personal data both within and outside the EEA while ensuring adequate protection of those personal data.

Unfortunately there is some divergence between Member States in the implementation of Article 26(2) provision. The UK Information Commissioner encourages Data Controllers to base exports on their own assessment of risk, while other countries require that all such exports receive prior approval from the Data Protection Authority. Given the number of identity providers and service providers involved in e-Research, it seems unlikely that Data Protection Authorities would want to receive requests to authorise every such transfer. As with the relationship between identity providers and service providers above, a common approach that scales effectively to large numbers of relationships is essential.

3.3 eduGAIN - Federated Access to Web Applications

eduGAIN is an infrastructure developed in the context of the GÉANT project to enable trustworthy exchange of information for authentication and authorisation purposes among the GÉANT partners and other cooperating parties.

eduGAIN has been designed to address inter-federation, to enable users from one federation to access services provided by another federation. This approach requires an infrastructure that supports the exchange of information between different entities (often located in different countries), a legal framework (such as a contractual agreement) and Data Protection Directives that ensure that the users' personal data are securely handled.

The eduGAIN service offers a solution for Web Single Sign-on (WebSSO), which enables users to log in to multiple services, provided by different federations, using a single, one-

step log-in process. There is a strong demand to extend the SSO to other applications and services areas. A typical example is a researcher who needs access to Grid based services and scientific instruments that do not use web browser clients and protocols. Development is ongoing in the eduGAIN team to support these use cases.

The picture below provides a high-level overview of the eduGAIN model, illustrating the basic eduGAIN components and their relationship. It is important to note that eduGAIN builds on top of national IDFs and that different IDFs can chose which of their entities can take part in the eduGAIN service.

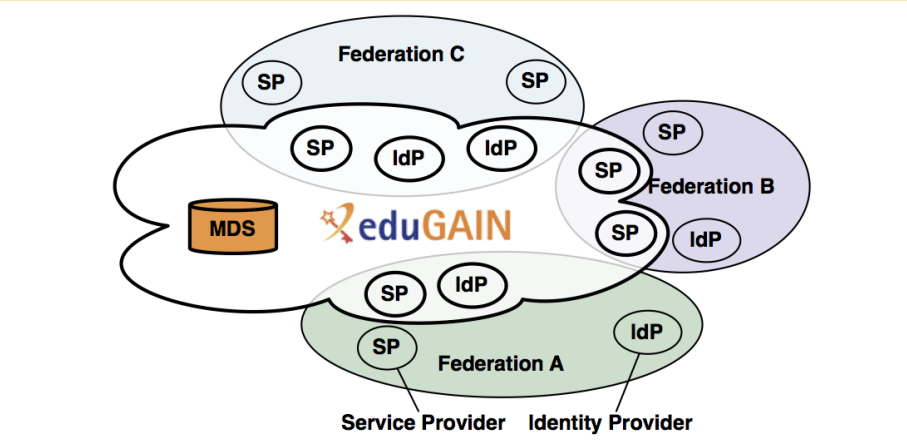


Figure 1: eduGAIN technical architecture (courtesy of eduGAIN)

An area that needs addressing by the eduGAIN team relates to transferring Personal Data needed by services to authorise users, which typically means across national borders. This exchange of users' information is regulated under the Data Protection Directive as well as by national laws. The eduGAIN team, together with the REFEDS community, is working to implement a scalable solution to provide the relevant information to resource providers to enable the service delivery without violating the mentioned above data protection laws. More information about the proposed model and the motivations for that can be found online.

The table below provides a summary of eduGAIN's main features; it is important to note that eduGAIN builds on the existing IDFs and as such it inherits most of the features described in the previous section.

eduGAIN Strengths	eduGAIN Challenges
Infrastructure built on standard technologies to address the inter-federation requirement.	Extend eduGAIN technology and protocols to support non-web applications
Privacy aware environment complying to the EU Data Protection Law.	Motivating entities involved in eduGAIN to meet data protection adherence
Addresses many important use-cases for educational and research community as identified in the context of the GÉANT Project.	Extending membership and involvement from non-NREN federations.
Limited requirements for national federations to participate.	Requirement for harmonisation of higher Levels of Assurance via eduGAIN.
Inclusion of eduGAIN metadata provides federated access to services offered by the participating federations.	Participating federations required to modify their metadata management practices to match inter-federation agreement.
Single identity from the home organisation is used to access a range of the federated resources and applications.	Increase the number of services available via eduGAIN, by engaging more with relying parties.

Table 2: eduGAIN overview

3.4 eduroam - Federated Access to the Network

The aim of eduroam is to enable federated access to the network: users with valid eduroam credentials can get online on any eduroam network in the world. eduroam is available in Europe, Canada, USA and Asia-Pacific region (Australia, Japan, Korea, New Zealand and others).

eduroam, which in 2012 celebrates its 10th anniversary, is the most successful example of a federated infrastructure used in the academic community (and in some cases extended to other communities). The eduroam infrastructure is built upon two main technologies: the IEEE 802.1X authentication standard, to securely handle users' credentials and a hierarchy of RADIUS proxy servers, to transport users' credentials.

The picture below depicts how eduroam works in the case in which a user coming from University B (unib.nl) in the Netherlands, tries to get connected to the eduroam network of University A, still in the Netherlands. Upon successful authentication of the user, which takes place at user's home university (University B), the user can get online.

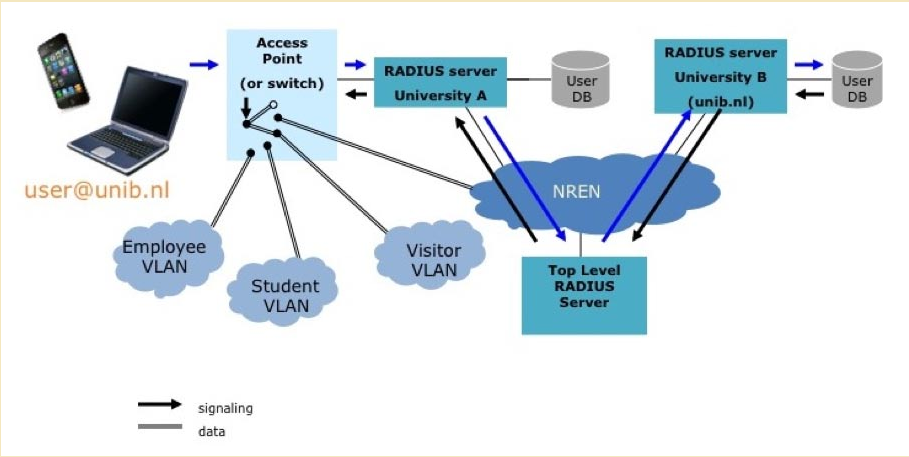


Figure 2: eduroam architecture (courtesy of SURFnet)

The hierarchical model followed by eduroam, mimicking the DNS hierarchy, has issues with domains that do not fit into this model, like ".org" or ".eu" or ".edu" domains that are used by some organisations. RADIUS over TLS, an IETF standard since May 2012 offers a solution to this problem. The new features in RADIUS over TLS allow the use of a more dynamic trust model, where connections can be established between the users' home institutions and the visited institutions.

The eduroam service in Europe, which operates in the context of the GÉANT project, has evolved into a confederation: a federation of federations, where each country in Europe operates and it is responsible for their national eduroam operations. A European eduroam Operational Team is in place to ensure that international roaming can take place. Collaboration and coordination with other countries has initially been rather informal; because of the wider deployment and growing use of eduroam, the community has requested a firmer basis for eduroam governance worldwide. The Global eduroam Governance Committee was constituted in November 2010 and at the moment comprises seven senior representatives for North America, Asia-Pacific and Europe as well as a TERENA appointed independent expert; TERENA operates the secretariat.

Table 3 summarises eduroam's main features:

eduroam Strengths	eduroam Challenges
Infrastructure to offer secure access to wireless (and wired) networks based on well established and stable protocols (802.1X and RADIUS).	Extending beyond the educational domain and NREN and universities federations.
Privacy preserving technology (users' personal information is not forwarded to the visiting institution).	Addressing fine-grained authorisation requirements.
Offers access to all eduroam networks with the same credentials (no need to request additional credentials when moving to another institution).	Deploying solution to address the current limitations inherent with RADIUS.
Scales to a large number of connected institutions and works on all kind of portable devices.	Improving the consistency of the user experience when configuring eduroam on different devices.

Table 3: eduroam overview

3.5 The Moonshot Project

Project Moonshot is a JANET (the UK NREN) originating project that aims to address solutions which enhance Identity Federations as they operate today. Moonshot aims to provide a solution to the following problems:

1. **Support for Non-Web Applications**
   Even though Web browser provides a de facto interface to the majority of Internet services, many applications are either not web-based or are more effectively used through a native application. Examples are Outlook access to an IMAP server, Shell access to High-Performance Computing clusters or access to a chat service.

2. **Addressing scalability issues in discovering the Identity Provider of a user**
   Federations that nowadays have hundreds of Identity Providers are struggling to present the user a convenient interface for selecting his home organisation Identity Provider. If federations couple with other federations this problem only gets bigger.

3. **Support for multiple affiliations and federations**
   A user can at the same time be a student at one school and a teacher at another, can both be a teacher and a parent and can belong to the identity federation of the research network as well as to that of a professional or societal organisation he belongs to. Current trust fabrics have a hard time distinguishing these roles contacting the appropriate Identity Provider for assertions about the user.

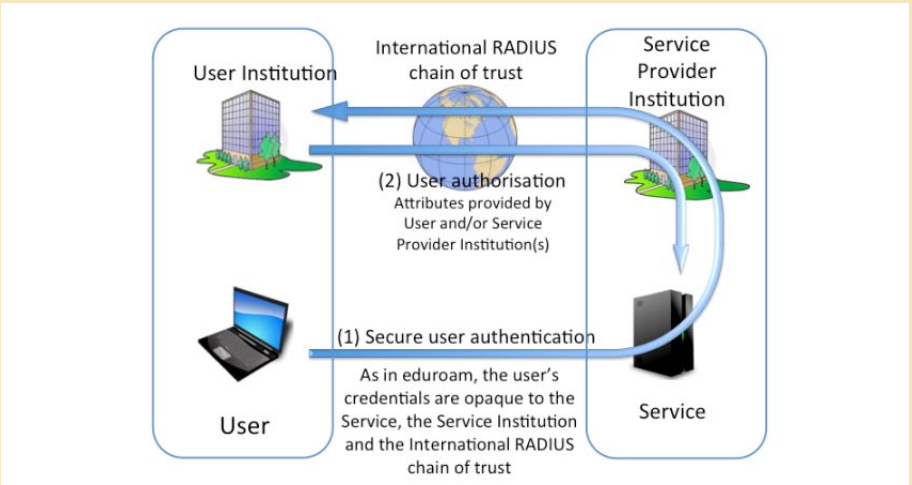


Figure 3: Moonshot Architecture

Project Moonshot addresses these issues by proposing an architecture (fig. 3) that builds on the foundational technologies of two successful federated identity infrastructures described above: Identity Federations and eduroam.

Project Moonshot combines the eduroam trust model (built using RADIUS servers), with EAP authentication standards for confidentiality of user credentials and SAML standards for the exchange of information about the user. In addition, a standard API (GSS-API) is used to provide a common interface to applications. This makes for a scalable solution that can be used for both Web- and non-Web applications. Project Moonshot also aims to provide solutions to the requirement for multiple affiliations and to define a model 'trust router' that performs a selection of potential trust 'paths' between Identity Provider and Relying Party.

The technology developed in the Project Moonshot is being standardised in the Abfab (Application Bridging for Federated Authentication Beyond web-sso) Working Group in the IETF. The table below provides a summary of Moonshot features.

Moonshot Strengths	Moonshot Challenges
Attempts to solve problems with accessing cross-domain non-web resources.	Requires updating standards on which it is based, what may create difficulties for wide industry adoption.
Builds on long-term AAI community experience and well defined goals.	Moving the communities forward with new developments is challenging.
Is an IETF proposed standard track activity. Benefits from the well established IETF standardisation process.	Needs development of new software to be widely adopted.

Table 4: Moonshot overview

3.6 Grid Infrastructures

The past decade has seen the emergence of an e-Science infrastructure in Europe encompassing resources from a large number of different providers and concurrently used by many different research communities. This e-Infrastructure, coordinated by amongst others EGI, is commonly referred to as 'the grid' – although the latter term is also used more widely to refer to any distributed infrastructure that combines resources from multiple organisations managed by different administrative domains. In EGI users are organised in virtual research communities (VRCs) and virtual organisations (VOs), which are logically distinct from resource providers.

The grid aims to coordinate the sharing of resources in a dynamic and multi-institutional setting to provide additional functionality beyond its constituent parts: brokering, workflow coordination, integration of compute and storage. In order for this to happen, interoperability and standards need to be defined at various levels: for resource access, for coordination and business logic, for data storage and management, network access and so forth. Given the intent of the infrastructure to span multiple organisational domains, two areas in particular have attracted attention:

- **Technology** – authentication, authorisation and accounting protocols, mechanisms to organise users in communities and express collective attributes like membership and roles, and a resource access methodology that emphasises local resource autonomy and independence;

- **Policy** – authentication assurance levels, identity vetting quality, traceability, acceptable use, and incident response.

The provisioning of collective services in the grid has been the driving force for one of the most characteristic aspects of the grid AAI: delegation of credentials to agents and

software services. In order to provision collective services, participating agents need to authenticate in order to deliver that service to the user. Given the potentially large number of agents involved, and a current lack of a single administrative domain managing those agents, these agents require credentials to authenticate amongst each other.

Underlying Technology

At the technical level, users and resources in the Grid today are identified using a public-key infrastructure (PKI), based on a trusted third party scheme of Certification Authorities (CAs).

The user is either explicitly issued with such a certificate, or can obtain one as-needed based on another form of authentication, e.g. by logging into a (academic) federation or by using an institutional account. The digital certificate binds the user or the resource 'name' to the data used to prove possession of this private key, thereby proving them a digital identity. The names thus issued are the key to access control (and data ownership) in the grid system and must be assumed to be unique within the scope of a grid.

In recent years, several such systems have been deployed by the research and education community; examples of these services are the Short Lived Credential Service (SLCS) developed by SWITCH, and TERENA's Certificate Service (TCS).

Although access to a resource can be based on explicit user lists, such an identity-based access control system does not scale. Grid software has therefore introduced VO-centred credential providers that can assert community membership, sub-groups, and roles of the user. The most-used mechanism for expressing VO membership is the Virtual Organisation Management System (VOMS), expressing VO attributes either as RFC3281 attribute certificates (ACs), or as SAML statements. A user would present both an identity certificate as well as a VOMS statement linked to this identity in order to gain access to the service, and the access control decision is then primarily based on the VOMS.

The user can now direct the service to perform a specific task, i.e. coordinate a workflow or compute and store the results elsewhere. In order for the service to perform these tasks, it has to be able to act in coordination with other services: at the end of a computational task, the resulting data set must stored elsewhere; or the broker will have to split the task in many individual jobs to be executed all across the grid. Except in some specific cases, the service should prove that any such actions originate with the actual user request. This is done using a *delegated* credential, carrying the verifiable identity and attributes of the original user.
The delegation method used most in the grid is an extension of the PKI model where the user can 'extend' the certificate path with 'proxy certificates' signed by the user credential, and thereby certify that a specific service can act on his or her behalf.
These proxy certificates carry the original identity of the user, but can in addition hold attributes and assertions by others, e.g. to hold credentials issued by a VO.

Trust Model

Collaborating across different domains can only work when the organisations involved can ensure the integrity of their own infrastructure, and their own policies and procedures are adhered to. This means establishing a policy and procedural framework around the AAI for authentication, traceability, incident response and accounting, and the production grid infrastructures both in Europe and elsewhere have established such a framework.

The European Policy Management Authority for grid authentication in e-Science (EUGridPMA), working with equivalent bodies in the Asia Pacific (APGridPMA) and the Americas (TAGPMA) in the International Grid Trust Federation (IGTF) at the global level,

provide persistent and unique naming of all users and resources in the grid. The assurance level is sufficiently high that the names can be used to determine long-term authorisation decisions, as well as to assign ownership of data in long-term storage archives. The minimum requirements for joining the IGTF as an identity provider are driven directly by the security needs of the participating grid infrastructures (relying parties or RPs), which include EGI.eu, and PRACE-RI in Europe, but also Open Science Grid, XSEDE, NAREGI, PRAGMA, as well as most national e-Infrastructures around the world. The RPs are directly represented as stakeholders in the trust federation and, jointly with identity providers, ensure the integrity of the trust fabric.

A policy framework which is largely common between all major infrastructures has been established to address community (VO) membership enrolment and management requirements, a harmonised incident response process, and a common end-user acceptable use policy (AUP). This common set of policies, which is fostered by the Security for Collaborating Infrastructures (SCI) group, allows users of the e-Infrastructure to enroll once and gain access to a global set of resources without the need to register and accept the policies and conditions of each and every participants. Grid resource providers accept the common AUP and VO policies as sufficient, relieving the users from the need to sign large sets of terms and conditions.

In authorising access to resources, the VO or user community plays a very specific role: it is the VO (and not the user's identity provider or 'home organisation') that determines the roles, memberships and rights on the user in a particular session. The authorisation decision at the site is therefore based on a combination of identity (for individual traceability) and VO attributes, coming from *different authoritative sources*.

With the general advancement in AAI technologies, the users' perception of certificate based access control changed significantly. More and more of the existing and potential new user communities look at certificate based access as a barrier to accessing resources, which is only partially addressed by the portal policy and current federated certificate systems.

EGI has recently undertaken a study on the use of traditional and new technologies for access control and identity management in Grid and on EGI. The EGI study concluded that federated identity management can facilitate access to grid resources for users, therefore motivating EGI movement towards broad adoption of federated identity management technologies and solutions within European research communities, although it also points at significant hurdles that still need to be taken. EGI intends to invest into the wider and more harmonised adoption of federated identity management systems on EGI and between NGIs.

As confirmed by the Federated Identity Workshop originating as an EIROforum initiative, there is common consensus to work towards increased use of Federated Identity Management (FIM) and federated services on Grid infrastructure, although there are both technical as well as policy issues still to be addressed.

Privacy Aspects

Minimum requirements for identity vetting in the IGTF federation today include face-to-face identity validation, and related controls address the need to provide globally-unique and long-term persistent names (subject identifiers), as well as traceability to individuals.

Sufficient information about the participants must be recorded and retained by the identity provider in order to allow for renewal of credentials. Since re-issuance of the identifiers to *assuredly the same individual* must be possible both after expiration of the certificate as well as following a revocation, this involves retention of identity information outside the digital domain.

Although this does imply a trade-off between user privacy and the resource provider's need for traceability (especially for data storage), regulatory requirements and the need to respond to incidents in an high-value, high-risk infrastructure have up to now excluded the use of anonymous or pseudonymous identifiers in the grid. Data *confidentiality* for end-users can be ensured by existing technical means, using distributed key management systems.

Grid AAI Strength	Grid AAI Challenges
Support for strong authentication (identity of users generally requires face-to-face vetting).	Managing large scale of digital certificates raises usability issues; approaches to improve the user's experience using digital certificates are being proposed.
Separation between authentication (based on the digital certificates) and authorisation (based on attributes controlled by VOs).	Ensuring that attributes maintained by different VOs are not community specific.
Grid infrastructure is cross-border and it can be considered as an example of inter-federation.	Leveraging identity federation infrastructures with grid AAIs (IDF to provide identities for the Grid users and for Grid infrastructures to consume that).

Table 5: Grid Infrastructures overview

3.7 PRACE: Access to European Supercomputing Facilities

PRACE, the Partnership for Advanced Computing in Europe, is a co-funded EC project which started in 2008. PRACE offers world class computing, data management resources and services open to all European public researchers. The need to maximise the usage of the facilities and to minimise the costs mandated a distributed Research Infrastructure, since no single site can host all the necessary systems required in terms of space, power, and cooling facilities.

PRACE has a mixed AA infrastructure. The basic service model enables users to have interactive access to sites on which they can run jobs. In this case, a users' access is handled on a per site basis and requires SSH (Secure Shell) credentials; each SSH credentials can only be used for the site for which they have been issued.

PRACE however also provides remote Single-Sign On features, implemented via the GSISSH protocol (the modified version of OpenSSH to support Grid Security Infrastructure) and the usage of personal certificates (X.509 certificates), which must be issued by CAs accredited by the IGTF. Beside GSISSH, the PRACE infrastructure include other services like GridFTP, UNICORE (Uniform Interface to Computing Resources) helpdesk and accounting. GSISSH and the other services enable users to run jobs on different systems and to transfer data between sites.

Information about users, needed for the operation of the integrated services (those available via the SSO), is managed in a shared LDAP-based repository, where each partner has only write access to their domain, but partners can access all information stored in the database for infrastructure management purposes. This approach defines the basis of the PRACE AAI federation: each partner relies and trust the information provided by the others.

In addition to basic account information, the repository contains further additional attributes for authorisation purposes; for instance it gives information on which systems the user can access to run jobs. The attribute information is based on IETF standard schemas and a PRACE-specific schema that includes attributes specific for the PRACE application domain. This means that when interoperating with other AAIs (for instance IDFs) the information received from those AAIs needs to be complemented by the attribute information from the PRACE AAI.

PRACE Strengths	PRACE Challenges
Builds on long-term Grid AAI experience and well defined goals.	Integration with AAIs requires specific PRACE attributes.
Requires strong authorisation before X.509 certificates are issued.	General usability issues related to the usage of personal certificates.

Table 6: PRACE overview

The Umbrella Project

PaNdata is developing an identity system, known as Umbrella, for the users of the European Neutron and Photon sources. Umbrella is a pilot authentication/authorisation infrastructure whose goal is to federates the local user management systems used in the photon community of the participating facilities. Umbrella supports more than 30000 visiting scientists that twice per year for a limited period of time perform their experiments using neutron/photon facilities. The management of these experiments is handled by the so called 'local user offices', a few people at each facility, who enable scientists to access the facilities (i.e by providing necessary support for registration, stockroom, computer accounts, storage space for experimental data and so on).

The demand for federated AAI in this community is triggered by an increased need to access remote services, especially remote data access and remote experiment access. Because of the highly competitive type of research, the AAI needs to offer support for confidentiality, fine-grained access control, identity uniqueness and persistence, as well as being user friendly.

The main feature for the Umbrella AAI architecture is that it uses only one IdP, which is used by the whole facilities to verify the identity of the user; the rest of the information related to the users (which are needed for authorisation purposes) is stored at the local user offices databases. The Umbrella IdP contains references to the location (local user offices databases) from where to retrieve the rest of the information of the users. The main function of this single IdP is to guarantee a unique user identification and therefore a unique and persistent user identifier; this identifier is then used across the whole facilities.

The communication of the various elements of Umbrella (the Umbrella central IdP and the resource providers) is based on Shibboleth/SAML2.

Implementation within the photon/neutron community is scheduled for early 2013.

3.8 Cloud Infrastructures

Cloud Computing technologies are becoming a common way of provisioning infrastructure services on-demand that may combine computing and storage facilities as well as dedicated network infrastructure. A number of commercial cloud providers offer computing services on-demand spanning from the possibility of running virtual machines to platforms for developing user applications.

There are two main type of cloud services: public clouds, typically operated by commercial companies (Amazon, Rackspace etc.) and private clouds, typical operated by a specific user group.

The main benefits of clouds are on-demand services with pay per-use that does not require user and organisations to own hardware or build their own infrastructure, and a possibility to dynamically scale resources required for solving specific tasks. The distributed character of cloud resources means that tasks and applications can run anywhere in the world depending on the physical spread of the cloud provider infrastructure.

Security Considerations in Cloud Infrastructures

Cloud technologies are based on hardware virtualisation, which allows for management of virtual computing resources (scaling, migration, reconfiguration) independently from the applications layer. In theory cloud based virtualised applications should run in the same way as non-virtualised applications; in reality in many cases moving applications to the clouds requires their redesigning to support dynamic deployment and configuration. Cloud based service virtualisation provides an additional level of security due to the separation of applications executing environments and the possibility of using preconfigured security enhanced virtual machines images.

A concern with public clouds relates to data security, as in most of the cases the data is stored 'in the cloud' on servers whose location is conceptually nebulous. Typical questions users ask are "Where are my data? Are they protected? What control has the Cloud Provider over data security and location?" and also "Who has access to my data? Is the usage statistics collected and how it can be used?"

The distributed nature of Cloud computing, where backup servers and data can be located potentially anywhere causes problems in understanding which of the national data protection laws apply to the cloud instance and its usage – something often defined in a click through the terms-of-use, which is often not fully comprehended by the end user.

Identity Management and Cloud Infrastructures

Clouds are becoming more and more popular among researchers as it allows them to quickly obtain necessary computing facilities when they are needed (and often without the additional procedures to access organisational or community Grid resources). Research organisations are also using more and more virtualisation platforms offered 'in the cloud', on top of which other more community or science specific infrastructure services and applications can be built. At the same time NRENs are defining strategies to offer cloud services, in some case contracted by commercial providers. In establishing these arrangements, NRENS are seeking a model that allows cloud services to be used without compromising the AAI arrangements already offered to researchers and research organisations.

In moving out into the cloud, researchers are also moving outside of the AAI that has grown up to support their workflow within the federated and grid spaces. Many cloud offerings are actually a step backwards in terms of access management and group management, with this being local to the commercial offering. Researchers may find themselves unable to make use of institutional credentials and unable to connect their entire research group to the cloud service without explicitly asking other users to sign up for a new service.

In order to optimise the benefits of cloud approaches, researchers and research organisations will need to review their identity management approaches and decide how much of the AAI infrastructure can be comfortably virtualised, outsourced and managed by third parties. Cloud providers are increasingly talking about Identity-as-a-service offered as part of the cloud package, and this will challenge all of the approaches to identity discussed in this report.
Cloud type identity services will emerge in two ways – via organisational identity services fully outsourced to cloud providers and through the already established use of social identities as credentials for a range of services.

The added complexities of identity management within a cloud environment have been recognised in the development of new standards to support such processes. The Simple Cloud Identity Management (SCIM) specification has been introduced to tackle specific workflow problems for cloud identity. SCIM does not specify any particular authentication

or authorisation schema, but instead specifies a way for a variety of known endpoints (directories, group management systems, required services) to be seamlessly and easily linked together to provide an AAI. In this sense, SCIM is addressing the provisioning side of the identity management workflow. SCIM is a very new specification, but as the identity space grows it is likely that this type of cross-walking approach to AAI may become more popular and indeed necessary.

In summary, the main issues that should be addressed to make cloud environment and cloud based infrastructures secure and trustworthy for wide range of scientific applications are:

- Standards to facilitate interoperability;
- Secure operation of cloud infrastructure in line with national data protection laws and directives (typically addressed by cloud providers);
- Clear and explicit terms of use and licensing for cloud services;
- Integration of cloud based infrastructure and access control services with existing AAI.

Cloud AAI Strengths	Cloud AAI Challenges
Ability to view AAI as a commodity service.	Need to address common concerns about data security – both stored data and personal data related to user identity.
Cloud based services virtualisation provides also an additional level of security by separation of applications executing environments and a possibility to use preconfigured security enhanced virtual machines images.	To exploit the potential benefits of clouds virtualisation, new trust and security management mechanisms need to be developed.
Virtualised approaches to cloud AAI potentially cost-efficient for organisations.	Architectures and models for access control and trust management in clouds still evolving.
Cloud providers can guarantee high level of availability, data recovery and security of their infrastructure and platform.	Need to balance with integration with campus infrastructure and legacy applications.
Major cloud service providers (such as GoogleApps, Amazon) either use or plan to implement SAML based federated access to user deployed infrastructure or applications.	Privacy issues related to the way these providers handle users data and/or stored data.

Table 7: Cloud AAI overview

3.9 Summary

The chapters above provided an overview and an analysis of the existing infrastructures used by the research and education community and the use-cases they address. Because of the diversity of the requirements coming from the various communities and because of some limitations within the current technologies, it is impossible to have a one-fits-all infrastructure. However some trends can be observed:

- All infrastructures evaluated provide Single-Sign-On for the users, although the technology used varies: SAML for Identity Federations, 802.1X + RADIUS for eduroam, X.509 certificates for PRACE and most of the eScience infrastructures.

- No single AA(A) technology can be universally adopted, but there should be mechanisms in place to allow for integration of different technologies. The current trend within the research network environment is to converge Grid identity infrastructures (based on X.509 certificates) and NREN-operated identity infrastructures based on RADIUS and SAML. There have been successful initiatives to leverage users credentials issued on campus (i.e. via Shibboleth) to issue grid certificates; and via Moonshot to implement SSO for different type of applications. Authorisation requires mechanisms for identity data aggregation for authorisation decisions; this is particularly challenging in an heterogeneous

environment and it is often implemented via complex systems. There seems to be consensus that service-oriented architectures can hide complexity while offering rich mechanisms, including better support for accounting.

- Enhancement to the current identity federations are needed to support the eScience requirements, such as stronger authentication vetting.

- Cloud computing is considered as a cost-effective solution for the data deluge problem; however there are still security considerations (how to maintain ownership of the data, under which legislation the data are stored etc) that need to be addressed. At the present, privacy can only be guaranteed by private clouds, which are mostly community specifics. In the research and education community, some NRENs are investing to offer private clouds.

  The lack of real standardisation means that each solution is vendor-specific; this makes the migration to a different provider rather complex.

- Consistent implementation and interpretation of the legal requirements in the Data Protection area is essential when building an international infrastructure.

4 Recommendations

4.1 Introduction

The purpose of this study has been to envisage what work is required to create a future proof AAI for SDI. In the previous chapters, the study identified different user-community requirements that would benefit from an SDI and the gaps in the existing infrastructures in supporting all identified requirements for authentication and authorisation infrastructure.

The study team took the view that:

1. Existing infrastructures are well established and they support the user-community requirements for which they have been designed.
2. There is not an existing AAI that can support all requirements.
3. Harmonisation of existing infrastructures should be pursued by bridging the gaps between the existing AAIs.

This chapter provides a set of recommendations addressed to different stakeholders, which will address the identified gaps.

4.2 The Vision

We can imagine that:

> ***" In ten years time, most research data are readily discoverable and the vast majority of data are open and in the public domain. Data are used ethically according to the norms of the research community, including fair attribution. "***

" In ten years time, most research data are readily discoverable and the vast majority of data are open and in the public domain. Data are used ethically according to the norms of the research community, including fair attribution."

Changes are needed to implement this vision; these changes should primarily address the technology which needs to evolve to cope with data deluge, data access and data availability in the longer term. The technology alone is however not sufficient to ensure the deployment and the sustainability of the SDI; funding bodies and existing organisations (NRENs, libraries, research centres) will have to evolve accordingly.

Historically, research data have been seen as the researcher's private property or perhaps as a national asset that should be protected or exploited as a commodity. There is now growing consensus in the research community and in the funding bodies of the value of open data, and there is a general movement toward more openness and less restriction in terms of data usage and availability.

To move toward a vision of wider data sharing and cooperation, trustworthy infrastructures will be an important condition; the successful future infrastructure will need to be able to manage a variety of access policies and positions where there are legitimate restrictions on data access to protect human privacy and cultural and natural heritage.

The trend for more open access to data has, however, to take into consideration the high levels of commitment and manpower dedicated to research, and has to recognise the fundamental drivers of competition, priority and IPR in the research community. The terms of grants given to researchers often dictate and drive the approach taken to these issues. The future SDI should therefore:

- be trusted by both data/information providers who are concerned about access control to data;
- be trusted by researchers who are concerned about the safety and confidentiality of their research results.

The recommendations below dictate a pragmatic approach to achieving a flexible AAI for SDI, building on and improving existing architectures, recognising the diversity of requirements in different research communities and addressing the need to support an open, yet protected, trustworthy infrastructure.

4.3 Technical Recommendations

Technologies to enable an AAI for SDI already exist and are maturing well. The core focus for technical development should be on enhancing existing infrastructures, supporting standardisation and interoperability work and embedding technologies consistently across EU countries.

1. Support the use and standardisation of federated technologies for network, service and application access across Europe (and beyond) to implement access control and identity management, particularly in geographical areas that have seen slow adoption. Specific support should be given to inter-federation to achieve cross-disciplinary and cross-boundary requirements and to create a common, but distributed AAI for SDI.

2. Develop and improve existing infrastructure in the following areas:
   a. tools to better enable AAI for mobile access and mobile devices.
   b. AAI schemas to support the uptake and use of persistent identifiers such as ORCID.
   c. tools to allow for effective accounting across the highly distributed, heterogeneous infrastructures envisaged for global research data.
   d. support for the use of social networking identities and groups as both identity providers and attribute providers for the SDI.

3. Aim to simplify the adoption level of existing infrastructures. User friendly solutions should be explored to provide simplified authorisation mechanisms, whilst offering rich internal tools for aggregation and authorisation decisions.

4. Leverage cloud virtualisation and complex infrastructure management technologies to build distributed dynamic and secure data processing environments beyond the traditional boundaries of institutions.

4.4 Policy and Practice Recommendations

There is growing realisation that existing policies and practices do not meet the needs of ever evolving technologies and more critically, are being interpreted and implemented differently across the EU member states.

1. Support the development of a common policy and trust framework for Identity Management. REFEDS (Research and Education FEDerations) the international body led by TERENA to coordinate Identity Federation processes, practices and

policies and to discuss ways to facilitate inter-federation work, should play a pivotal role to facilitate this process. REFEDS should evolve to become the equivalent of the IGTF role for eScience. Communication among different groups (EC, REFEDS, IGTF, ESFRI, libraries and policy maker) should be improved.

2. A closer collaboration between national federations, national eScience centres and libraries is needed to ensure that existing services are offered in a consistent way to the users. One way to improve this collaboration could be to organisationally aggregate NRENs and other types of infrastructures (i.e. grids/clouds), hence offering one interface to both network, grid and cloud services. This is the model followed in the Netherlands (and in other countries as well) where SARA (the Supercomputing facility) and SURFnet (the National Research and Education Network) are being moved under one umbrella (SURF).

3. Help communities to create AAI infrastructures using common tools and based on well documented and standardised practices and policies. Provide the necessary training to support this aim.

4. Utilise the current work on service provider classifications within federations to provide assurance that fundamental rights will not be over-ridden by the exchange of attributes necessary for federated access management.

5. Enable research projects with tools to check and enforce IPR policies, particularly when moving to cloud technologies that use infrastructure virtualisation and data processing facilities. Integrate such policy enforcement tools and infrastructure with geolocation and presence information.

4.5 Legal Recommendations

Developments in this area should focus on achieving clarity, consistency and user-friendly tools for implementation. The main law to be considered in this area is the European Data Protection Directive (95/46/EC) and its revision, the draft Data Protection Regulation (published in January 2012).

1. Extend the Legitimate Interests justification (Article 7f of the Data Protection Law) to cover international transfers, as proposed by the draft Data Protection Regulation to permit the use of a common legal framework for all e-research involving European researchers or services. It has also been suggested that the Regulation might be accompanied by a review of the current arrangements for export of personal data: including provisions suitable for use by overseas universities and public research organisations would further assist collaboration between European and overseas researchers.

2. Provide clarity about Consent and Legitimate Interest. The existing relationships (employment, site licences, etc.) between individuals, identity providers and services in education and research cast doubt on whether Consent (Article 7(a) of the Data Protection Directive) is the appropriate justification for processing in federated access management (the draft Regulation would make the use of Consent within an employment relationship even more questionable). Instead both identity providers and service providers appear to have a legitimate interest in providing access to the services their members seek to use, which justifies them exchanging information necessary to do so. Consent can then be reserved for information that is not necessary to provide the service, but where the user wishes to enhance the service by providing it.

3. Commission a study to investigate how adequate protection of personal data can be achieved by incorporating lightweight agreements into existing relationships

between researchers, projects, services and home organisations, whether these are user-mediated, organisation-mediated or (for example in VO-based authorisation) both.

4. Create and support a clear statement on the legal status of processing opaque identifiers, implemented consistently across Member States, to support the use of privacy-protecting identifiers in federated access management. This statement should offer the possibility for service providers to treat suitably protected opaque identifiers as non-personal data or, at least of representing a very low risk to privacy with correspondingly light regulatory requirements.

5. The EC should provide clear guidelines for the Member States on how to implement the Draft Data Protection Regulation when it enters into force. Fragmentation in the implementation of the current Data Protection Directive has caused several problems and as in fact hindered international collaboration, which is the cornerstone of research. The EC should, together with the Article 29 WP, organise training for the Member States representatives to avoid cultural interpretations of the laws.

6. Member State Data Protection Laws should be aligned with EC Data Protection Directives/Laws. This lack of clarity and the resulting variation in Member State interpretations on what is subject to personal data regulation make it difficult to exchange attributes between countries. If the same attribute is considered personal data in one country but not in another it is unclear whether the attribute can lawfully be transferred between them. Even within a country there may be significant problems: opaque identifiers are specifically designed to make it impossible to identify the real-world individual, yet classifying them as personal data appears to impose on data controllers a duty to do just that in order to satisfy subject access requests and breach notification requirements.

4.6 Recommendations for Funding Agencies, EC and Member States

1. Funding should be allocated to develop and enhance the convergence of eInfrastructures. Particularly EC funding should be directed, where possible, to consolidate and harmonise established systems rather than creating new ones. A proliferation of localised solutions unable to leverage existing infrastructures must be avoided. The deployment of an SDI can only be successful if Member States embrace it and provide the necessary funding to ensure that universities, libraries and research centres can connect to it.

2. EC must bootstrap plans to implement the Digital Agenda; strong measures should be taken to highlight the benefits of an integrated SDI and to engage with different stakeholders. The fragmentation of the current landscape and the different needs/interests of the stakeholders involved in different initiatives may lead to a situation that can hinder the actual deployment of an integrated SDIs. This situation can be avoided by ensuring participation of different players (both at technical, policy, national and international level) at the very beginning.

3. Periodical studies should be funded to assess the emergence of new technologies and the penetration of existing infrastructures at national level. The results of these studies can be used to inform recommendations, workshop or specific actions.

4. National federation operators should act to expand the coverage of their identity federations to include research centres and libraries. As highlighted in the FIM

paper there is an increased interest from the eScience communities (including the arts and humanities) in using federated access. Funding should be made available to expand the coverage of national identity federations to fully embrace these communities.

5. Member States should invest to implement Article 7(f) of the Data Protection Directive in a consistent way. Article (7) states that "processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection under Article 1".

6. The operation of data centres, storage and cloud facilities should be (whenever possible) entrusted to existing and specialised centres. The ability to provide long-term (measured in decades rather than years) storage and accessibility is key to the research, particularly in the modern society where data are available in electronic format. Whilst libraries have expertise in curating the data, in many cases they have no expertise in operating data centres or storage facilities. Libraries should not operate data/storage centres, but should instead rely on infrastructures offered by (inter)national data centres moving towards a model in which the management of the content is decoupled from the management of the infrastructure.