Federated Identity Management for Scientific Collaborations

Paper Type: Research paper

Date of this draft: 25th February 2012

Abstract

Federated identity management (FIM) is an arrangement that can be made among multiple organisations that lets subscribers use the same identification data to obtain access to the secured resources of all organisations in the group. Identity federation offers economic advantages, as well as convenience, to organisations and their users. For example, multiple institutions can share a single application, with resultant cost savings and consolidation of resources. In order for FIM to be effective, the partners must have a sense of mutual trust.

A number of laboratories including national and regional research organizations, are facing the challenge of a deluge of scientific data that needs to be accessed by expanding user bases in dynamic collaborations that cross organisational and national boundaries.

Driven by these needs, representatives from a variety of research communities, including photon/neutron facilities, social science & humanities, high-energy physics, atmospheric science, bioinformatics and fusion energy, have come together to discuss how to address these issues with the objective to define a common policy and trust framework for Identity Management based on existing structures, federations and technologies.

This paper will describe the needs of the research communities, the status of the activities in the FIM domain and highlight specific use cases. The common vision for FIM across these communities will be presented as well the key stages of the roadmap and a set of recommendations intended to ensure its implementation.

Keywords
federated identity management, security, authentication, authorization, collaboration, community

1. Introduction

Federated identity management (FIM) is an arrangement that can be made among multiple organisations that lets subscribers use the same identification data to obtain access to the secured resources of all organisations in the group. Identity federation offers economic advantages, as well as convenience, to organisations and their users. For example, multiple institutions can share a single application, with resultant cost savings and consolidation of resources. In order for FIM to be effective, the partners must have a sense of mutual trust.

A number of laboratories including national and regional research organisations, are facing the challenge of a deluge of scientific data that needs to be accessed by expanding user bases in dynamic collaborations that cross organisational and national boundaries. Many of the users have accounts at several research organisations and will need to use services provided by yet more organisations involved in scientific collaborations. All these identities and services need to be able work together without the users' being obliged to remember a growing number of accounts and passwords. As the user communities served by these organizations are growing they are also becoming younger and this younger generation has little tolerance for artificial barriers, many being the relics of technology and policies that could, if reasoned, also evolve. This "Facebook" generation [1] has triggered a change in the attitude towards IT tools. One expects to be able to share data, software, results, thoughts and emotions with whom they choose, when they choose. The boundaries between work and social life are less sharp, and it is expected that tools blend into this environment seamlessly. The interaction with commercial services such as the social networks must not imply that the users and scientific communities relinquish control over access to resources and security polices. The frequency of use will vary between the different users. Some will use these new tools continuously each day while others will log in a few times per year. This implies that operation has to be very intuitive, preferentially in a style known from common commercial devices and applications (PCs, smart phones, tablets etc).

Driven by these needs, representatives from a variety of research communities, including researchers from European photon/neutron facilities, social science & humanities, High-energy physics, atmospheric science, bioinformatics and fusion energy, have come together to discuss how to address these issues with the objective to define a common policy and trust framework for Identity Management and secure access to data based on existing structures, federations and technologies.

Many of these research communities are linked to the ESFRI [2] supported Research Infrastructure projects and the discussions have included national and international infrastructures that provide identity related services, standards forums and those whose responsibility it is to decide what identity mechanisms will be recommended to policy making bodies. While this activity is focused on Europe, the scientific communities have global needs and so interoperation with identity management systems in USA and Asia are essential.

The discussions have been promoted via a series of workshops on Federated Identity Systems for Scientific Collaborations. The first workshop was held at CERN in June 2011 [3], the second at RAL in November 2011 [4] and third in Taipei in February 2012 [5]. As a result of these workshops, a common vision for FIM across the scientific collaborations has emerged along with the desire to see this implemented with a roadmap and a set of recommendations.

2. The need for Federated Identity Management

The idea of organising a workshop on identity management came from a discussion with IT leaders of EIROforum [6] laboratories during an EIROforum IT working group meeting hosted by ESA in Frascati, Italy in January 2011. That meeting showed that these laboratories, as well as national and regional research organizations, are facing similar challenges. Given the potential scope and impact of this subject, the EIROforum members agreed to widen the participation and take an inclusive approach.
The grouping represented by the participants to this series of workshops has an informal mandate to achieve the objective of defining a common policy and trust framework for Identity Management based on existing structures, federations and technologies.

3. Scientific Communities

The participants to the series of workshops included representatives from a variety of scientific communities; many of them linked to the ESFRI supported Research Infrastructure projects. Each of the participating communities is described below. The workshops are open to representatives from all scientific communities that are facing similar challenges. Each community has identified a contact/architect that has helped organise the workshops and gathered material and opinions from within their community.

3.1 High Energy physics

CERN, the European Organization for Nuclear Research, is exploring fundamental physics, finding out what the universe is made of and how it works. The instruments used at CERN are particle accelerators and detectors. Accelerators like the Large Hadron Collider (LHC) boost beams of particles to high energies before they are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions. CERN is not an isolated laboratory, but rather a focus for an extensive community of scientists that spans more than 60 countries. These scientists typically work remotely from universities and national laboratories in their home countries. CERN leads the World-wide LHC Computing Grid project (WLCG), which is a global collaboration linking hundreds of computer centres worldwide. It was launched in 2001 to provide a global computing resource to store, distribute and analyse the data generated by the LHC. The infrastructure, built by integrating thousands of computers and storage systems, enables a collaborative computing environment on a scale never seen before. The WLCG serves a community of more than 10,000 physicists around the world with near real-time access to LHC data, and the power to process it.

The High Energy Physics (HEP) community is already successfully using federated identity and single sign-on for WLCG in the form of x509 certificates accredited by the of the International Grid Trust Foundation (IGTF [7]). The use of the research and education general federations and the associated TERENA Certificate Service

(TCS [8]) is likely to grow in importance if and when national CAs decide to move to TCS. In most HEP distributed infrastructures, a high level of trust and fined grained authorization are required and this is current role for x509. Support for non-browser-based applications is also need for which Shibboleth is often used, while access to traditional or simple services (like Wikis or Web portals) would benefit from the simplicity and ubiquity of OpenID. Credential translation services are being implemented to convert the credentials to different formats and are seen as a possible way of integrating these different technologies. For example, the EMI Security Token Service [9] and project Moonshot [10] are both developing such translation services. However, for support, cost, ease of deployment and maintenance purposes, limiting the number of credential translation services and planning for a careful integration within existing services is essential. An option would be to rely on a central service to issue x509 credentials for Grid-specific services, instead of relying on home-issued x509 credentials. Identity Providers (IdPs) could interface with such a service via a credential translation service. Several services already exist, for example the TERENA Certificate Service and the CILogon Service [11]. Security issues linked with the use of federated identities are also expected, and would need to be addressed:

- Used in the context of the Web browser, when a user is attempting to access a service, identity federation implementations typically prompt the user for credentials either via a sudden redirection to the IdP or via a popup window. In most cases, the user does not type the URL of the IdP. This behaviour creates an ideal opportunity for phishing attacks. Phishing attacks are extremely difficult to prevent and are the main cause of compromised accounts at most HEP sites.
- The trust in the federated infrastructure relies on the fact that both service providers and IdPs are trusted. If one of the participants becomes malicious, then other participants might be directly affected. For example, a malicious service provider could pull (and misuse) as many attributes as possible for all the known users at a given IdP. The implications of operating with malicious participants needs to be evaluated and understood.

The goal would be for HEP sites to enable their IdP(s) to be fully accredited by the IGTF for both x509 and non-x509 authentication systems. As such the IGTF-accredited IdPs would be considered as trusted IdPs, performing more rigorous vetting to provide an elevated level of assurance (LoA) and as a consequence grant a level of trust proportional to the LoA of the credentials presented by the user. HEP sites would also benefit from the more general and gradual adoption of identity federation in the community. In particular, external users/visitors would no longer need to have a local account to access resources at a site they are visiting while local users would be able to access resources at collaborating institutes using their usual account. As a result of these changes, it is believed that the operational and support costs associated with managing user accounts would be reduced.
Such an identity federation in the High Energy Physics (HEP) community would rely on:

- A well-defined framework to ensure sufficient trust and security among the different IdPs and relying parties, Anyone can indeed setup an IdP and "offer" to authenticate users from a given HEP site, or attempt to pull attributes from that HEP site, irrespectively of their intent or trust relationship with the site. The IGTF is already providing several authentication profiles and some regulation [12] enabling identity providers to become accredited and therefore trusted IdPs. All the CAs used in WLCG (100+) are for example accredited.
- Clear LoAs are needed, should an organisation decide to open access to its resources to external users. For example, credentials issued by an IdP who verifies the real identity based on a government-issued document and affiliation of its users are more trustworthy than Facebook or Google OpenID credentials. The level of authorization within the different services at a given organisation should also depend on the LoA of the credentials presented by the user.

In order to achieve this, the following areas need to be worked on:

- The existing policy set would need to be modified to integrate the concept of identity federation.
- Attributes definition, management, integrity and release would need to be addressed. Issues are to be expected with regards to sensitive attribute release. For example, some services require the "nationality" attribute to be presented, while some federations are more conservative and will not even release the real name of the user.
- Security procedures and incident response would need to be reviewed. Today, each resource provider is for example responsible to terminate the access of known compromised identities. With identity federation, this responsibility will be shifted to the IdP though resources providers will insist on the ability to revoke access.

3.2 Life Sciences

ELIXIR is an ESFRI research infrastructure that is building a secure, evolving platform for biological data collection, storage and management, consisting of an interlinked set of core and specialist resources. One of the

main roles for ELIXIR will be to enable the linking of biomolecular data to biomedical and clinical data. This will necessitate the creation of data resources that contain data that is restricted to certain individuals for ethical, societal, legal or other reasons – so-called ELSI data. An example of such a system is the European Genome-phenome Archive (EGA) which is currently maintained and deployed by EMBL-EBI [13]. It contains about 150 large data sets. Access to these is managed by approximately fifty data access committees (DACs). EMBL-EBI does not itself grant access, but implements the access mechanisms that are required by the particular DAC in question.

As ELIXIR matures there will be many systems such as EGA that will provide secure access to authorized users of the data contained within them. Biomedical data service providers (SPs) running the systems therefore will need to implement authentication, authorization and auditing procedures. The deployed FIM system should deliver solutions that can be used by the biomedical service providers to make the access control layers required for the data. The doubts among SP's have arise from the fact that FIMs currently do not cover all countries in Europe or globally. Biological research community is by far the largest - over three million scientists in Europe alone, and especially smaller SP's cannot support multiple ways of authentication. Technical deployment of FIM is seen challenging by some SPs and they require support. A step forward would be introducing biomedical data service providers with the expertise of the existing European networks of trust, and demonstrate the impacts of integration of FIM with key biomedical SP use cases.

ELIXIR will put in place a series of pilot studies to investigate different aspects of the platform. The purpose of the first pilot study will be to provide the option for users to authenticate themselves with federated identity management when applying access to ELSI data.
The purpose of the second pilot study will be to create an electronic workflow that will automate the procedure by which a researcher whose identity and affiliation is confirmed with FIM applies for and gains access to a particular dataset in the system through an appropriate Data Access Committee (DAC).
The pilot studies will identify requirements that are mandatory for a DAC to grant access to a dataset. In particular, we aim to approach the Nordic Centre of Excellence in Disease Genetics (NCoEDG) DACs to authorize access to the data made available through the EGA. The Finnish Haka identity federation service provides technology, experience and required standards to manage federated identities in both projects. We will also identify those requirements that must be addressed to allow infrastructure from both pilots to interact automatically with the Nordic (Kalmar Union) and the European (eduGAIN) networks of trust.

We seek to demonstrate that administrational costs are not increased as the number of accounts continues to increase in the future, and it is possible to identify the host institution and role of the user with the aid of FIM.

We also hope to demonstrate that the use of FIM will increase the security of access. We think that this is very likely as there is already considerable evidence that users are less likely to share their usernames and passwords with other users when these will also give access to institutional resources such as private email.

FIM is also expected to allow automatic data access expiry when the account holder departs from his home organization and his accounts are closed. In the event that EGA took part in the pilot then they would not affect the current EGA authorization mechanisms but would allow Haka users to use their home organisation's identity to authenticate to EGA services.

The second pilot builds a demonstrator for data access application/authorization platform that could be integrated to the EGA user management and data dissemination workflows. It will allow separate views for the applicant, DAC and EGA to manage and audit data access authorization process in an automatic manner.

Both pilots have a synergy where FIM provides a reliable and automatic way of identifying data access applicants as part of a particular host institute. Workflow also provides an audit trail for the process.

3.3 Humanities

In the Humanities there is a number of research infrastructure projects (see the list below) that have identified the need for an AA infrastructure that offers SSO and single user identity for individuals and scales with an increasing number of users. Although initially some projects have entertained the thought of creating a single user database, currently all are now pursuing or planning this through use of FIM. Specific for the Humanities domain is the sometimes-sparse distribution of users over very many academic and educational institutions that

makes it difficult for those infrastructures to directly influence the setup and functioning of the user home institute identity providers. However the user home institutions are normally always easily connected to their respective national IDFs.

Currently the humanity infrastructure projects does not have a complete AA infrastructure in place. Furthest along the road is CLARIN that has created a (slowly) growing CLARIN Service Provider Federation (CLARIN SPF) that is making contracts with the national Identity Federations and is exchanging the CLARIN SPF metadata with them. The CLARIN SPF requires user attributes ePPN or ePTID, although the **recommended** set of attributes also comprises three other attributes: "cn", "mail" and "o" or "schacHomeOrganization". Although the technology has been proven to work, there are a few serious problems in rolling it out and expanding to the foreseen user domain. This is caused by reasons that differ across participating countries.

1. Many identity providers do not expose user-attributes even if the national IDF requires them to do so. This can be caused by erroneous software configuration but more often because the IdP management is afraid for being responsible for privacy law violations.
2. Some IDFs require the IdP administrators to explicitly give permission to pass on the user-attributes to any SP. This policy makes it difficult to easily give large number of users access CLARIN SPs. Such an "opt-in" policy does not scale well.
3. The research infrastructures mentioned are EU wide; however as an IDF is not available in every participating country it requires the use of a special identity provider for the 'homeless'. Maintaining such a special identity provider is an extra effort.

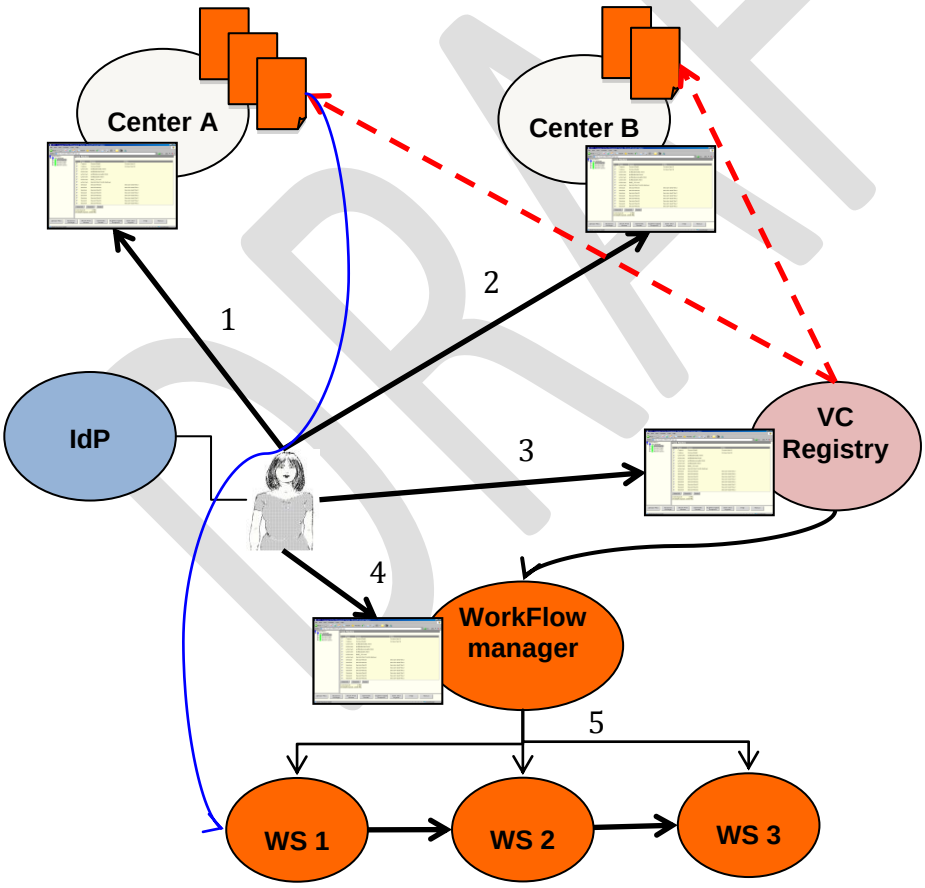

Figure 1 Humanities Use Case

The Figure 1 above describes a use case from the CLARIN project with the creation and processing of a so-called Virtual Collection (VC). A VC is a collection that is created by a user for a specific purpose by selecting resources from other collections and storing their references with the metadata for this VC. The steps involved include:

1. user selects suitable resources at center A using a specific app at center A after logging in via his organizational account
2. user selects suitable resources at center B using a center specific app making use of SSO
3. references are added to a Virtual Collection registry via a VC registry app for future reference and use
4. The VC is processed by a workflow of LT Web services
5. The identity of the user is delegated to shielded WSs that can use it to access resources.

It is not yet clear what technology can be best used to achieve step 5. Currently CLARIN is investigating in cooperation with the Dutch national Grid initiative the use of OAuth2.

Currently the following projects within the Humanities are known to use or work towards using a Federated Identity Management Infrastructure.

CLARIN http://www.clarin.eu/.
CLARIN (ESFRI) has aimed from the beginning at using FIM as an AAI solution. Currently there is a CLARIN Service Provider Federation of currently ten service providers that has signed contracts with national IDFs (currently SURFfed, DFNAAI, HAKA and more to come)

DARIAH (ESFRI) http://www.dariah.eu/
Although no results have been realized yet, DARIAH plans to establish a Shibboleth based federation across Europe, ideally eduGain and it wants also to investigate integration of user-centred approaches as openId. Experiences were gained using X.509 (robot- and short-lived) certificates obtained via Shibboleth login for job submission purposes in the TextGrid DARIAH sub-project.

CESSDA (ESFRI) http://www.cessda.org/
No experiences or plans using FIM. But the cooperation with CLARIN and DARIAH in the DASISH project will also put CESSDA on the track of using FIM.

DASISH http://www.dasish.eu/
DASISH is an ESFRI cluster project for the humanities infrastructure projects: CESSDA, CLARIN, DASISH, ESS and SHARE. For solving AAI issues, it will rely on the available expertise and experiences gained in CLARIN.

Project Bamboo http://www.projectbamboo.org/
A privately funded multi-institutional, interdisciplinary project of humanities scholars, librarians, and information technologists aimed at advancing arts and humanities research by using shared technology services. Project Bamboo has several authentication use-cases based on FIM.

3.4 European Neutron and Photon Facilities

There are every year more than 30’000 visiting researchers performing experiments at the about two dozen European photon and neutron facilities. In most cases, these facilities are part of national research institutions. These users conduct their research in many disciplines, ranging from materials sciences, life sciences, physics, chemistry to environmental sciences, and even studies in archaeology and cultural heritage. The largest single scientific user community at these facilities is structural biology. Practically all protein structures, which are the basis for e.g. drug developments, are determined at these facilities. Research at these facilities is predominantly performed by academic institutions, but there is also a significant commercial - mainly pharmaceutical - activity. This already indicates that confidentiality in accessing scientific data and / or metadata plays a significant role in a federated data infrastructure.

Compared to a standard laboratory X-ray source, the brightness (photon intensity on a small spot) of a modern synchrotron light source is higher by 12 (!) orders of magnitude. This allows measurements to be performed, which are impossible with a conventional laboratory source. The current demand for data storage resulting from these experiments varies strongly and represents megabytes for some types of experiments and up to terabytes for others like structural biology, lens-less imaging or tomography measurements.

There are, however, two recent developments which boost the demand for storage space significantly.

- Initially as by-product from the detector development for LHC, large-area two-dimensional pixel detectors have been developed for the keV photon range, which have revolutionized the measurements for imaging techniques, especially tomography and structural biology. These detectors exhibit excellent resolution paired with extremely high signal to noise properties. The price to pay is that the data size is increased by orders of magnitude.
- Free-Electron Laser (FEL) facilities produce light another 10 orders of magnitude higher as compared to synchrotrons. Facilities in Germany (European XFEL), Japan (SCSS) and Switzerland (SwissFEL) are under development, the US facility (LCLS) is already in operation. Experience from LCLS shows that there is a strong increase in the data volumes collected.

In contrast to other research fields, e.g. HEP, users of photon / neutron facilities are very mobile. Experimental setups are more standardized and because of the high overbooking researchers are used to go to those facilities, which offer the best conditions for a specific experiment. According to a recent study [14], more than 30% of photon facility users and more than 20% of neutron facility users perform experiments at more than one facility.

The facilities are constantly entering new generations. Instruments and experiments are becoming more and more complex and sophisticated. The increasing complexity is accompanied by more complex and time-consuming analysis.

These challenges, as well as best-practice considerations make it increasingly impractical to leave data archival and curation to individual researchers. Sharing data within or across collaborations and scientific communities is becoming a major issue. These activities, however, should as much as possible be under the control of the users.

To address these challenges and to enable the creation of a federated, open data infrastructure for the users of the neutron and photon sources, almost all major European facilities are collaborating in the PaNdata and/or CRISP projects.

As mentioned above, confidentiality between researchers and between facilities plays an important role.

Up to now, there is generally no established procedure for data access across facilities. Users, who want to combine data sets from different facilities, have to access these facilities individually, using different locally established procedures and locally provided tools. At each facility a user possesses a dedicated account and accounts of the same user at different facilities are not linked to another. Thus, the lack of a common identity management makes it difficult to introduce common services and has a significant negative impact on standardisation of services among facilities.

Access to the individual facilities is controlled locally by the respective local Web-based User Office (WUOs). According to international practice, a user has access to experimental data if he/she has participated in the respective experiment and is part of the corresponding proposal. To be able to perform an experiment at a facility a scientists needs to go through a number of steps. There are slight differences from facility to facility, but the general workflow is usually very similar.

1. The first step is always registration of the responsible scientist and all members of the collaboration, who intend to participate in an experiment, at the facilities Web-based User Office (WUO). This step has to be made at each facility individually.
2. The facility maintains a user databases and verifies that the user is unique, and has not registered earlier under a different account name. This step has currently to be made at each facility individually.
3. The collaboration submits a proposal describing experiment and resources requested. Currently, this needs to be done per facility.
4. At each facility proposal review committees are installed, which select at regular intervals the best proposals and assign beamtime to them.
5. If the proposal is successful, the users come to the facility (typically for a few days) and perform their experiment. The results are published in standard scientific journals.

As mentioned above, confidentiality between researchers and between facilities plays an important role, access control to any experiment or user information is therefore fine grained.

For handling all the administrative issues, each facility runs a user office with several persons and a WUO as the IT tool. In Europe, these WUOs are based currently on two systems: DUO-type (PSI, HZB, DESY, SOLEIL,

FRM II[♣], MAX-lab) and SMIS-type (ESRF, DIAMOND, ANKA). New facilities will likely adopt one of them. An essential part of each system is the user database. Currently, there is no connection between the various databases of the different facilities and users have to register individually at each facility with all the related administrative overhead.

This situation leads to the following requirements:

- In view of the dramatic developments in accelerator and detector technology, resulting in a steep increase in data production, a new IT environment is urgently required.
- This environment must be based on unique EU-wide user identification with SSO capabilities.
- The system has to respect the confidentiality interests of users and facilities.
- All laboratories have invested considerable effort to cater for the local environment (e.g. security and accommodation aspects). These investments must be preserved.
- The new system will not replace the existing local WUOs, but only add new functionalities.
- The new identification system must be able to run in parallel to the existing local systems.
- The administrative load should be not higher (but indeed lower) than with the existing WUOs.
- In view of the number of facilities under construction, planned or in discussions the number of European users is not expected to increase by more than a factor of two compared to its present value (30'000+), even if the definition boundaries will be weakened (e.g. for remote users). Thus, in view of the well-established structuring of the researcher community according to facilities, scalability of a FIM solution is not an issue.
- The system must allow for efficient bridging to other scientific domains.

This led to the Umbrella system [9], which has been developed as prototype in WP2 of EuroFEL and taken over within the PaNdata and CRISP FP7 projects. The basic concept has been defined and is currently (Spring 2012) tested in a 'friendly user' phase. The full system, however, will still be under development for many years to come.

Conducting experiments at the large photon / neutron facilities is very attractive, which leads to a typical overbooking of the order of 2-3, hence the chance for an experimental proposal to be accepted is correspondingly low. On the other hand, and in order to promote scientific innovation, the threshold for new users submitting proposals must be as low as possible. This means, that the user account management (registration etc) as basic administrative action must be as simple as possible. On the other hand, user identification is also the basis for user access to sensitive procedures (access to facility, accelerator and experiment areas, to experiment data etc). On top of that, all confidentiality aspects and requirements between users, but also between facilities have to be respected. This leads to a hybrid identification system, as foreseen within the Umbrella system, with one central identity provider and minimal central information just sufficient for user identification and with the other personal, authentication and authorization information located at the local WUOs. Depending on the type of application, the identification strength required ranges from a low Google-handshake-type level to a high level (in-person identification at a facility user office) for enabling access to sensitive information.

Because of the high overbooking, teams will ask for beamtime a several facilities, which will increase the probability that data for an experimental investigation is scattered over several facilities. Because of the high data volumes from the multi-dimensional detectors it will often render it impractical to transfer raw data to the home institutes of the users. Instead, the trend will go to the installation of virtual analysis centres, where large experimental datasets will be stored and remotely analyzed. Such centres may be community owned or commercial cloud systems. All these options, however, will require standardized remote access to experimental data and computing resources. In comparison to web-based tools, remote data analysis will require much stronger authentication, authorisation and auditing mechanisms to be implemented (multi-level federated identity management).

An important issue in this context will be confidentiality. There is strong competition between the different researcher teams and also between the facilities hosting these teams. The situation is even more complicated, because the composition of these researcher teams is not stable but rather variable with time (patchwork collaborations), where e.g. postdocs (often being the most active members of the teams) are migrating between

[♣] development status

different teams. In view of the fact that each facility deals with hundreds to thousands of proposals per year there must be an automatic system, in order to keep the administrative load at an acceptable level. That means that, a system has to be developed, which (a) uniquely indentifies a user and (b) in a temporally fine-grained way determines, if a specific user has access to a specific dataset. As the Principal Investigators (PIs) of an experiment are the only ones, who have detailed knowledge about the participation of each researcher in their team, they will have to be part of the data access management by determining the composition of their team in the user database. As a consequence the authorization management which is currently typically in hands of service providers must be extended and based on scalable procedures where certain tasks are delegated to the members of user community.

In principle, the elements for solving this important issue exist. (a) the Umbrella system has the information for identifying users and (b) the local WUOs have all the necessary information for linking data sets to proposals. The goal, therefore, is to combine these elements in order to provide an efficient tool for controlling remote access to data at the participating facilities.

FIM in general and Umbrella in particular are central topics in WP 3 of PaNdata ODI and WP 16 of CRISP. This issue is the basic element of many other issues raised in these EU programs. This is especially true for remote data access. Another hot topic is remote experiment access. For some experiments, in part due to the boost in brightness, the duration or a single experiment is so short, that it does not justify a travel from the home institute to the facility. Consequently, remote access to an experiment – both passive (read online spectra) as active (remote control) becomes increasingly attractive. Social media will in future become more important. And, as probably the most important single topic, proposal procedures should be harmonized on the European scale. To accomplish this is mainly a political issue, as one is potentially touching some of the autonomy of the individual facilities. Reporting is only vertically and there is no political incentive for coordinating facility operation on a transnational scale. On the other hand, a harmonization on that scale will immediately foster cooperation between facilities in areas, which are not accessible nowadays.

A specialty of experiments at photon / neutron facilities is that travel / accommodation support of researchers is not covered by grants from national research programs but is handled to a large extent by the respective facilities, which are again supported via specific EU programs. As one is talking here about 4-digit number of visiting scientists with on the average two visits annually this implies a significant organisational load for the facilities. There are specific EU programs for photon (CALIPSO) and neutron (NMI3) facilities. Here, FIM plays obviously an important role and merging Umbrella with the local web-based user office (WUO) systems is well under way without a need for modifying established work flows.

In addition to facility-oriented support there is increasingly also a trend for community-oriented support. The most prominent example is BioStruct X for the structural biology community. For this type of research, in addition to the measurements performed at the large facilities there are other experimental steps (e.g. crystallisation, imaging measurements), which are in part performed at other facility or university laboratories and the goal of BioStruct X is, to perform an over-all evaluation of all aspects of an experimental study. This leads to a more complex experiment admission work flow. Definitively, negative impacts on facility friendliness (separate user databases) and user friendliness (parallel proposal review processes) have to be avoided, but in ongoing discussions a satisfactory solution based on the Umbrella concept is emerging.

3.5 Climate Science

In the climate science community and the wider field of the environmental sciences, researchers use data from numerous sources. The ability to combine such data is critical to the understanding of the Earth system. At the same time, integration of such datasets can be difficult especially if as is often the case, they are distributed across a range of data providers. Management of access across multiple domains, identity management and data licensing are barriers to access and consequently many valuable datasets are under used. One particular activity in climate science, CMIP5 because of its large scale and international scope has formed a focus to tackle many of these issues. The Coupled Model Comparison Project, Phase 5 is an internationally coordinated set of climate model experiments organized under the auspices of the UN World Climate Research Programme. Model data has been generated from super computer simulations from around 50 participating modeling centres around the world. The size of the prospective data archive – approximately 2.5PB – and the challenge of distributing this data to the user community necessitated a federated access solution.

The Earth System Grid Federation (ESGF) was developed and deployed to meet the requirements of CMIP5. This included a complete FIM solution. The system uses dual mechanisms for single sign-on: OpenID for

browser based interactions and MyProxyCA to issue short-lived user credentials for thick clients. SAML was used for both attribute management and authorization interfaces. Access control is required to audit access and enforce a simple set of licence terms. Hence, only a low level of assurance is required for user authentication.

ESGF is a distinct federation with IdPs independent of existing national providers. This has made it possible to adopt an agile approach to the system and to tailor to the needs of the user community. Nevertheless, there is a need to integrate with existing national and international FIM infrastructures. EGI-INSPIRE is one project focused on this goal. Credentials can be translated or exchanged between domains (ESGF and EGI) using for example the Security Token Service from WS-Trust. However, there is a need to ensure associated levels of assurance are communicated and also the provenance of credentials to ensure appropriate access is associated with a given set of credentials. The IGTF and other similar bodies have an important role to play in standardizing inter-federation access.

Given the low level of security needed for access to the CMIP5 archive, there is a danger that access control can be perceived as an unnecessary burden to members of the user community. However, the access control infrastructure is essential if as intended the ESGF system is to be more widely applicable across the Earth science community. Given the range of datasets and application scenarios there are likely to be use cases requiring much higher levels of security.

ESGF has followed a flexible approach to federation membership but there is a need for policies and agreed SLAs between members to ensure operational issues are dealt with effectively into the future.

In the UK, a new initiative CEMS presents new use cases for federated identity management. CEMS is the facility for Climate and Environmental Monitoring from Space, a public private partnership to host services and data to support climate change research and the development of commercial downstream applications. As a collaborative venture, FIM will be essential to the infrastructure that is developed. CEMS will use cloud technologies to exploit the benefits of co-locating services and applications next to large data volumes. The dynamic provision of resources is a key characteristic of clouds to exploit in order to respond to changes in demand for resources. This by its nature presents particular challenges to address both in terms of trust and confidentiality and the consequent level of security it is possible to achieve. Where PKI has traditionally used a very static model of trust, these use cases will demand the dynamic creation of credentials and trust domains to manage the provision of virtual networks, storage and processing resources.

Relevant projects include, CEMS, ESGF and CMIP5, Metafor, IS-ENES, CORDEX, Exarch, Climate Data Exchange, GENESI-DEC.

3.6 Other relevant parties

The work of the European E-infrastructure Forum, which published a report on the requirements for Pan-European e-infrastructure resources and facilities [18], concluded that, based on the information gathered from 28 ESFRI projects, consistent identity management and single sign-on is a fundamental requirement. As such the work described in this paper is of potentially far wider impact that initial set of scientific communities represented at the FIM workshops.

Representatives of national and international infrastructures that provide identity related services, standards forums and those whose responsibility it is to decide what identity mechanisms will be recommended to policy making bodies.

This included the services provided by the pan-European HTC and HPC e-infrastructures EGI [19], DEISA [20] and PRACE [21] as well as the Open Science Grid [22] in the USA. The EMI [23] project has outlined its plans for identity management within its future Grid middleware releases. GEANT [24] presented the eduGAIN service for exchanging trustworthy identity related information between partner federations as well as plans for the upcoming Moonshot project. TERENA [25] presented its Certificate Service for servers, users and software.

The International Grid Trust Federation (IGTF) is a body to establish common policies and guidelines between its Policy Management Authorities (PMAs) members and to ensure compliance to this Federation Document amongst the participating PMAs. The IGTF does not provide identity assertions but instead ensures that within the scope of the IGTF charter the assertions issued by accredited authorities of any of its member PMAs meet or

exceed an authentication profile relevant to the accredited authority. The IGTF maintains a list of trust anchors, root certificates and related meta-information for all the accredited authorities, i.e., those that meet or exceed the criteria mentioned in the Authentication Profiles. The Distribution contains Certificate Revocation List (CRL) locations, contact information, and signing policies. The IGTF constituency consists of our three member PMAs: the APGridPMA covering Asia and the Pacific, the EUGridPMA covering Europe, the Middle East and Africa, and The Americas Grid PMA covering Latin America, the Caribbean and North America. All registered members in each regional PMA are also members of the IGTF. These include identity providers, CAs, and their major Relying Parties, such as the international Grid Deployment and Infrastructure projects.

The Trans-European Research and Education Networking Association, TERENA, supports the work of REFEDS (Research and Education FEDerations) which aims to exchange Identity Federation processes, practices and policies and to discuss ways to facilitate inter-federation work. Its participants represent national identity federations and many come from National Research and Education Networks ('NRENS').

4. Analysis of common needs

The characteristics and current status of the user communities present have been captured in Figure 2 below. The user community column identifies the scientific research community represented; the other projects column lists those projects which are working with the user community and have an interest in identity management; the #users column gives an estimate of the number of individuals in the user community and hence an impression of possible scale of usage; the chosen technology column shows which technologies are in use or being considered by the community; the status column indicates the level of maturity of identity management deployment; the IGTF column indicates if the community is making use of the International Grid Trust Federation's policies, procedures or services.

user community	other projects	# users	chosen technology	status	IGTF
photon/neutron facilities	EUROFEL, PanData, CRISP	>30,000 visiting researchers per year	Shibboleth/SAML	Umbrella prototype	no
Social Sciences and Humanities	DARIAH, CLARIN, CESSDA, DASISH, Bamboo	hundreds now, potential for 10000+ across SSH	Shibboleth/SAML	CLARIN SP federation - will see if they can use eduGAIN	yes
high energy physics	WLCG	10,000+ globally	X509	production	yes
earth sciences	Federation, GENESI-DEC, CMIP5, Metafor, IS-ENES, CORDEX, Exarch, Climate Data Exchange	5000+ for CIMP5	OpenID, X.509 and SAML	production - earth system grid	not yet but foresee for EGI integration
life sciences	ELIXIR, BioMedBridges, BBMRI, NCoEDG & potentially 10 BMI ESFRI projects	3 million researchers access data via EBI website each year	no chosen yet	security included in BioMedBridges project workplan	no

Figure 2 user community characteristics and status

Distilling the requirements presented the session concluded there are many common needs and hence scope for agreement. In particular, commonality is most evident in the following domains:

- Single Sign-On access control
- Ease of use for part-time users (many researchers only access the ICT systems concerned in-frequently or on a part-time basis)
- Controlling access to data sets or repositories is the most foreseen of use of identity management across

the user communities
- Support for citizen scientists and researchers without FIM facilities (i.e. homeless having no suitable hosting institute or organisation) users is essential
- There is a wide range of tools and technologies already deployed
- A smooth transition from existing systems to a federated identity management system is essential

5. A Common Vision for Federated Identity Management

Based on the requirements and constraints identified by the scientific communities described above, we have identified the need for a common policy and trust framework for Identity Management based on existing structures and federations either presently in use by or available to the communities.

This common policy and trust framework needs to support the following:

- **Multiple technologies with translators including dynamic issue of credentials**. No single technology can meet the need of all communities. Translators between one type and another will be required to allow credentials from one community to be used on other services and this translation will often need to be dynamic.
- **Implementations based on open standards and sustainable with compatible licenses.** These are essential for interoperability and sustainability.
- **Different Levels of Assurance with provenance.** A single Level of Assurance in the quality of authentication cannot meet the need of all communities. Credentials issued under different levels will need to include the provenance of the level under which it was issued.
- **Authorisation under community and/or facility control.** The assignment of attributes to individual users within a given community for use in authorisation decisions needs to be managed by that community. Federated IdPs cannot fulfil this role.
- **Browser & non-browser federated access.** The wide-range of applications in use in the various communities include many which do not have a simple web-browser front-end.
- **Well defined semantically harmonised attributes.** For interoperable authorisation across many service providers it is necessary for the names and possible values of attributes to be well understood and standardised. This may be very difficult to achieve between different scientific communities.
- **Flexible and scalable IdP attribute release policy.** Different communities and indeed SPs within a community are likely to require a different set of attributes from the IdPs. The IdP policy related to the release of user attributes and the negotiation mechanism needs to be able to provide this flexibility. Bi-lateral negotiations between all SPs and all IdPs is not a scalable solution.
- **Privacy and data protection to be addressed with community-wide individual identities.** There are many use-cases identified which will require the release of personal data to identify individual users. Clearly this has to be managed in a way that satisfies all legal requirements for data protection.
- **Attributes must be able to cross national borders.** Many of the scientific use cases require user attributes from an IdP in one country to be used by an SP in another country. Data protection considerations must allow this to happen.
- **Attribute aggregation for authorisation**. Attributes will need to be aggregated from different sources of authority including federated IdPs and community-based attribute authorities.
- **Easy integration with local service provider (SP) environment.** The ease with which federated authentication and the related authorisation services can be used is an important consideration in the design of any new system. SPs are likely to want to support multiple means of authentication.
- In terms of privacy and security, the common policy and trust framework also needs to meet **specific requirements** from some communities, such as Biomedical, where competition between different research groups requires scoping within a given trust context.

There are various essential operational and usability issues that need to be addressed in our common framework. These include:

- **Risk Analysis**. This needs to include consideration of the use of an identity federation from the point of view of the community infrastructure. The implications of having a malicious SP in a federation, for example, need to be considered. This risk analysis will help prioritise the efforts needed to deal with various risks that have already been identified.
- **Traceability.** Identifying the cause of any security incident is essential for containment of its impact and to help prevent re-occurrence. The audit trail needs to include the federated IdPs.
- Appropriate **Security Incident Response** policies and procedures are required which need to include all

IdPs and SPs.

- **User friendliness** of the framework must be addressed with the aim of lowering the barriers to users and providing transparency about policies, e.g. the "what" and "why" of identity management.
- The **Reliability and Resilience** of the framework services are essential usability issues that must be addressed.

A number of legal, policy and trust issues must also be addressed including:

- **Contracts or SLAs** between communities and federations. These agreements should be developed in a scalable way, so that the maximum number of participants can be included. Bi-lateral agreements between many different communities and many identity federations may be difficult to achieve.
- One attractive way of addressing scalability could be to define **Standards of Trust** or codes of conduct similar to the Authentication profiles and guidelines developed and maintained by the IGTF.

6. Recommendations

List of Recommendations (need to be specific, what do we want, and who do we want to do it)
A point that was raised at the workshop was to somehow prioritise the recommendations – so we should think about the order in which they have to happen.

User friendliness
As already mentioned in the introduction, the attitude of users towards FIM tools has changed. They should be simple and intuitive and fit to the many other IT tools used in daily life.

Bridging
It will not be possible to provide one tool, which fulfils all demands. FIM is important and will be even more important in many scientific, commercial and social applications. Therefore, bridging between the various communities is a central issue with an efficient mapping of the respective attributes. Here, again user friendliness is an issue with the goal of maximum transparency and with requiring minimum actions by the users of these systems.

Flexibility
The field of identification techniques is vividly developing, e.g. in the government administration and banking sectors. New tools should be developed keeping this in mind and being ready to adapt to new technical approaches if appropriate.

Attributes release
There is a need to balance usability against user privacy and confidentiality. It has become evident through the workshops that there is considerable inertia weighted against this from the legal considerations and financial risk. Schemes to bag attributes into policies of varying trust level and then mapping service providers to these could work. It would need both new middleware deployed and policy agreement with IdPs concerned.

Attributes harmonisation and scalability
It is difficult for a project spanning multiple national infrastructures to get consistent attribute information. Reaching agreement across all the stakeholders will be difficult. Again middleware to broker consistent views might be a workable approach.

Levels of Security
A *one size fits all* model for levels of security supported for a given FIM system will not scale into the future. This is exemplified in the Biomedical community where there is a wide range of security levels needed but more widely, it is essential if single FIM systems are re-used across multiple domains and communities. We need to promote work on the standardisation efforts for levels of assurance and communication and enforcement of LoA within systems and across boundaries from one to another. This is not easily achievable given the broad scope but nevertheless will be of increasing importance as projects and research activities span multiple domains e.g. inter-federation across national boundaries.

Administration of Policies
The focus of the workshop has been on policy and we have tried to stay technology neutral. However, it is

recognised that technology which can simplify the administration of policies for IdPs will definitely contribute to the acceptance and uptake of FIM systems.

Risk Analysis
A pragmatic risk analysis of the use of identity federation from the point of view of an infrastructure provider will be necessary to reassure the security officers at participating sites. For example, the implications of having a malicious SP in a federation need to be understood and incident procedures developed. Such a risk analysis should prioritise the various risks and hence focus available effort.

Funding model and governance structure
In order to implement this common policy and trust framework for Identity Management an agreed funding model is required with an appropriate governance structure.

Pilot Studies
Life Sciences service providers (SPs) need to be better informed on federated identity management services. This is especially true for SPs who provide data service which needs access control. For life sciences and medical research we plan to run pilot studies under the auspices of ELIXIR and/or BioMedBridges. We anticipate that these will allow us to explore the requirements in more detail, which we already know are very diverse, as well as serving to alert more potential stakeholders to FIM. It is hoped that this will generate increased interest amongst this community as well as better defined requirements. This approach could be adopted across all the communities by working with the same set of FIM system providers and each community exploring their own use cases and requirements.

7. Relevance to European Policy

The e-Infrastructure Reflection Group (e-IRG) published a white paper in 2011[26] that includes a section on authentication and authorisation infrastructures (AAI). It outlines objectives which are consistent with those presented by the user communities in this paper:

"*The overall objective is to establish and maintain the level of mutual trust amongst users and service providers that is needed for an open ecosystem to function. As an e-Infrastructure matures and its user community grows, requirements for aligning authentication and authorisations grow as well. This must translate into:*

- *Improved usability, lowering the threshold for researchers to use the services.*
- *Improved security and accountability, which often conflicts with the usability requirement.*
- *Leveraging of existing identification systems, such as that of the employing organisation.*
- *Enhanced sharing, allowing willing users to minimise the burden of policy enforcement.*
- *Reduced management costs, freeing resources for other service or research activities, and providing a sound basis for accounting.*
- *Improved alliance with the commercial Internet, which also improves interaction between scientists and society.*"

The eIRG white paper also recommends the integration of different identity technologies and the development of a roadmap. The contents of this paper provide a concrete example of how these recommendations are being implemented.

8. Next Steps

What are the next steps we will perform in order to develop our roadmap and ensure our recommendations are implemented.

- Endorsement of the common vision paper by the scientific communities (March-June2012)
- Fourth workshop for confirmation of endorsement by the scientific communities, Max-Planck Institute for Psycholinguistics in Nijmegen, Netherlands on 21, 22 June 2012
- Interaction with other bodies

9. References

[1] Facebook generation http://www.wisegeek.com/what-is-the-facebook-generation.htm
[2] European Strategy Forum on Research Infrastructures http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[3] First workshop on Federated Identity Systems for Scientific Collaborations, CERN, June 2011, https://indico.cern.ch/event/129364.
[4] Second workshop on Federated Identity Systems for Scientific Collaborations, RAL, November 2011, https://indico.cern.ch/event/157486.
[5] Third workshop on Federated Identity Systems for Scientific Collaborations, Taipei, February 2012, http://event.twgrid.org/isgc2012/index.html
[6] EIROforum is a collaboration between eight European intergovernmental scientific research organisations that are responsible for infrastructures and laboratories: CERN, EFDA-JET, EMBL, ESA, ESO, ESRF, European XFEL and ILL, http://www.eiroforum.org/
[7] http://www.igtf.net
[8] http://www.terena.org/activities/tcs/
[9] http://www.eu-emi.eu/security
[10] http://www.project-moonshot.org/
[11] http://www.cilogon.org/
[12] https://www.eugridpma.org/guidelines
[13] https://www.ebi.ac.uk/ega/
[14] PaN-Data, Deliverable D2.3-rev, Policy Framework for User Data, R. Dimper 31-Jul-2011, http://www.pan-data.eu/imagesGHD/7/7e/PaN-data-D2-3.pdf
[15] EuroFEL, Deliverable D2.5, The Umbrella System, the Prototype Web-based Access Point, B. Abt, H.J. Weyer, 28-Apr-2011
[16] BBMRI http://www.bbmri.eu
[17] NCoEDG http://www.ncoedg.org
[18] European E-Infrastructure Forum, "ESFRI project requirements for Pan-European e-infrastructure resources and facilities", April 2010, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/eef-report.pdf
[19] http://www.egi.eu/
[20] http://www.deisa.eu/
[21] http://www.prace-project.eu/
[22] http://www.opensciencegrid.org/
[23] http://www.eu-emi.eu/
[24] http://www.geant.net/
[25] http://www.terena.org/
[26] eIRG White Paper 2011 http://www.e-irg.eu/images/stories/e_irg_whitepaper_and_comments_2011.zip

10. Acknowledgements

The work of Philip Kershaw is funded by the UK Natural Environment Research Council.
The authors would like to thank the following people for contributing to the material presented in this paper: Frank Schlünzen, DESY, Rudolf Dimper, ESRF, Krzysztof Wrona, EU XFEL

11. Biographies

Daan Broeder is deputy head of the "Language Archive" a unit of the Max-Planck Institute for Psycholinguistics and is responsible for the group developing the LAT archiving software. He was one of the technical coordinators in the CLARIN EU project and is a member of the CLARIN NL Executive Board.

Bob Jones is head of the CERN openlab project (http://openlab.cern.ch) which facilitates collaboration between CERN and its industrial partners to study and develop data-intensive solutions for scientists working at the next-generation Large Hadron Collider (LHC). His experience in the distributed computing arena includes mandates as the technical director and then project director of the European Commission co-financed EGEE projects (2004-2010 http://www.eu-egee.org), which established and operated a production Grid facility for e-Science spanning 300 sites across 48 countries for more than 12,000 researchers.

David Kelsey is Head of Particle Physics Computing at the STFC Rutherford Appleton Laboratory. He has held security related responsibilities within various Grids (GridPP, WLCG, EGEE and EGI), starting with the creation of the EU Certification Authorities Coordination Group in 2001. This subsequently resulted in the formation of the EUGridPMA and the International Grid Trust Federation (IGTF). Today he continues to lead the

development of security policy for both EGI and WLCG and represents these infrastructures inside the IGTF.

Philip Kershaw is a senior developer with CEDA, the Centre for Environmental Data Archival at RAL Space, STFC Rutherford Appleton Laboratory in the UK. He is a specialist in federated identity management and has contributed to the security architecture for a number of distributed systems including the Earth System Grid Federation and Contrail, an EU Framework 7 project to develop a system to support federated cloud infrastructures. He authored and contributed to a number of papers and abstracts in the area of federated identity management and access control for applications in the environmental sciences domain.

Stefan Lüders, PhD, graduated from the Swiss Federal Institute of Technology in Zurich and joined CERN in 2002. Being initially developer of a common safety system used in all four experiments at the Large Hadron Collider, he gathered expertise in cyber-security issues of control systems. Consequently in 2004, he took over responsibilities in securing CERN's accelerator and infrastructure control systems against cyber-threats. Subsequently, he joined the CERN Computer Security Incident Response Team and is today heading this team as CERN's Computer Security Officer with the mandate to coordinate all aspects of CERN's computer security --- office computing security, computer centre security, Grid computing security and control system security --- whilst maintaining CERN's academic environment and taking into account CERN's operational needs. Dr. Lueders has presented on these topics at many different occasions to international bodies, governments, and companies, and published several articles.

Andrew Lyall, Ph.D., ELIXIR Project Manager (http://www.elixir-europe.org/), European Bioinformatics Institute (http://www.ebi.ac.uk/). ELIXIR aims to create a sustainable infrastructure for biological information in Europe, laying the foundations for the impending biological revolution. Before coming to EMBL-EBI Andrew spent 15 years working in industry, primarily at GlaxoSmithKline, where he rose to the position of Department Head. He has also worked in the biotechnology sector, at Oxford Glycosciences and as a founding director of Confirmant Ltd. Prior to this he worked as an academic researcher at the University of Bristol and the Edinburgh University as well as the Royal College of Surgeons in Ireland. He read biochemistry and computer science at Imperial College and received his PhD in bioinformatics from Edinburgh University.

Dr. Tommi Nyrönen works as a development manager at CSC – the IT Center for Science Ltd. He is the Finnish contact in ELIXIR - European Life Science Infrastructure for Biological Information since 2007, and scientific representative in the ELIXIR interim board. After receiving Ph.D. in biochemistry (biocomputing) in 2000, he worked to develop computational drug design environment at CSC and co-founded a bioinformatics company FBD Ltd. He became a manager of CSC's bioinformatics services in 2003. He has been involved in a number of research and development projects building ICT services for science both nationally and at the EU level. His scholarly work covers structural bioinformatics and cheminformatics, scientific software, operating ICT services, and user training. He is an inventor in three medicinal patent families and an adjunct professor of computational drug design in the University of Helsinki, and a board member in the national graduate school for informational and structural biology.

Romain Wartel is the security officer for the Worldwide LHC Computing Grid. He has been involved in the operational security and policy aspects of several national and international Grid projects. He has for example led the operational security coordination team of the EGEE infrastructure between 2006 and 2010. He currently focuses on international security incident response and on improving the collaboration on security issues between different computing infrastructures.

Heinz J Weyer, PhD is member of the SwissFEL team and teaching at Basel University. He was scientific coordinator of the Swiss Light Source SLS and leading author of the Digital User Office DUO in use at most European Photon and Neutron large facilities. He is active in several EU projects for the development of the new generation of IT resources for the users at these facilities.