How many users does EGI have?

Cyril Lorphelin

Computing Center of IN2P3/CNRS

43 bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

E-mail: `cyril.lorphelin@cc.in2p3.fr`

Nuno Loureiro-Ferreira, Gergely Sipos, Tiziana Ferrari

EGI.eu

Science Park 140, 1098 XG Amsterdam, Netherlands

E-mails: `nuno.ferreira@egi.eu, gergely.sipos@egi.eu, tiziana.ferrari@egi.eu`

To build up a comprehensive understanding about the success, uptake and impact of the European Grid Infrastructure (EGI), one needs to look at several metrics. The number of EGI users is one of the most relevant metrics in this respect. While the importance of this metric is apparent, obtaining the number of users is quite a challenge due to the federated and distributed nature of the European Grid Infrastructure and of its user communities, as well as the existence of different mechanisms for users to access the resources. The Operations Portal is the most important tool to collect user statistics in EGI. The Portal generates statistics based on membership records obtained from Virtual Organisations. This paper presents the recent developments on the Operations Portal fostered by EGI-InSPIRE project members, about new reporting mechanisms put in place in order to improve reliability of the workflow by which EGI obtains user statistics.

1. Introduction

The European Grid Infrastructure (EGI) is the result of pioneering work that has, over the last decade, built a collaborative production infrastructure of uniform services through the federation of national resource providers that support multi-disciplinary science across Europe and around the world. An ecosystem of national and European funding agencies, research communities, technology providers, technology integrators, operation and support centres, over 350 resource centres, coordinating bodies and other functions has now emerged to serve over 21,000 researchers in their intensive data analysis. The EGI ecosystem provides a common foundation upon which the digital ERA can be brought online by building upon the strengths that have been developed over the last decade of deploying, operating, monitoring, accounting and supporting virtual research environments through a federation of national resource providers for multiple research communities at a European scale.

The EGI-InSPIRE project is a four year long EC-funded project that started in May 2010 to support the transition of EGI from a project-based system to a sustainable pan-European e-infrastructure. The project established a set of services to the overall EGI community, namely human services (for coordination and community building), technical services (for supporting the interaction of user communities), and infrastructure services (for securely accessing resources hosted by different organizations). These services are provided by EGI.eu (the coordination institute of EGI and EGI-InSPIRE), National Grid Infrastructures (NGIs), European Intergovernmental Research Organisations (EIROs), and by other various national and European projects that participate in the EGI collaboration.

Impact metrics, such as the number of users, are one of the most relevant indicators to evaluate the uptake and success of an e-infrastructure. Due to the distributed nature and complexity of the infrastructure, its access interfaces, and of its user communities, the number of EGI users is very difficult to capture. The EGI Operations Portal [1], developed and hosted by the IN2P3 Computing Centre in France, is the most important tool used by EGI to assess its user base. The Portal is used by different actors of EGI, namely resource centre administrators and managers of research teams to create and manage Virtual Organisations (VOs) on the production infrastructure. An EGI VO is a set of resources that are allocated from the grid infrastructure for a group of people who share similar interests, have similar goals, and need to work collaboratively using services of a grid. The Operations Portal provides a registry and categorisation of EGI VOs, as well as collects information about VO users from the 'Virtual Organisation Management Services' (VOMSes).

We have recently identified a few issues with the VOMS based membership collection process of the EGI Operations Portal. Issues like incorrect or inaccessible VOMSes, and the spread of new types of access methods outside of VOs control, lead EGI-InSPIRE project to re-evaluate the complete workflow and tools by which the number of EGI users is collected. The paper introduces the identified issues and the solutions that our teams already put in place or should put in place in order to address all of the problems.

The paper is organized as follows. Section 2 describes the Operations Portal. Section 3 describes the method by which the number of EGI users is obtained by Operations Portal.

Section 4 describes the issues of the method, while section 5 provides a description of the solutions that our team put in place to improve and extend it. Section 6 provides an analysis of the open issues, and finally section 7 describes the future challenges concerning user number calculation.

2. The Operations Portal

2.1 Overview

The Operations Portal is based on the 'actor's view' principle where each actor of the community has access to information accordingly to his role in EGI, such as grid operators (monitor the status of resources and grid services), researchers (make use of the grid for their work), and VO/site/NGI managers (provide services for specific groups). The Operations Portal provides various services to these actors, such as the Virtual Organization administration tool, communication tools (the broadcast and downtime system), and dashboards (operations, security, VO operations).

The Operations Portal has been designed in a modular way for maximal efficiency and robustness. The architecture consists of three parts (see also **Figure 1):**

1. A database – to store information related to users and VOs;
2. A web module – a graphical user interface integrated into the Symfony framework;
3. Lavoisier module – a data aggregation and unification service [2].

The Operations Portal has functional dependencies on a number of tools and middleware services to gather information about the grid and its users:

- VOMS servers;
- EGI Helpdesk (User support);
- Information system;
- Monitoring systems: Nagios , Pakiti , Gstat;
- VO database;
- Site Database;
- EGI Single Sign On system (SSO system);

The information is pulled from the external sources and cached in the Lavoisier module. This component controls the storage, consolidation and the 'feed' of data into the web application. By providing a consistent flow of information from various sources to the web application front end, it increases the robustness of the framework in case of problems while the retrieval of fresh information from the sources is not possible. The Lavoisier application was developed to enable easy and efficient cross-data source queries, independently of used technologies. Data views are represented as XML documents and the query language is XSL.

The web portal is structured around Lavoisier, which provides a unified view of data collected from multiple heterogeneous sources, and deals with:

- Retrieval of information coming from external sources and in different formats;
- Sorting, organization and cross-querying of information;
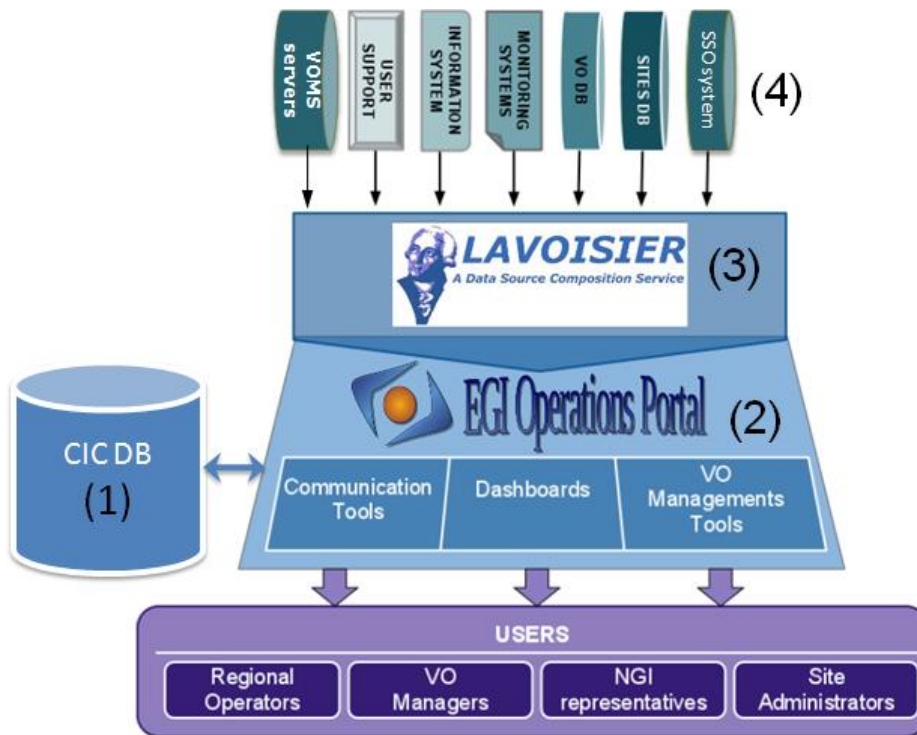- Unified access to information in a standard format and via web interfaces.

Figure 1 – Operations Portal architecture. The three components of the framework are (1) Operations Portal database, (2) Web module, and (3) the Lavoisier data aggregation and unification service. The information displayed by the portal is retrieved from several static and dynamic sources (4).

2.2 VO identity card

The Operations Portal records the life cycle of EGI VOs, and ensures the proper communication between VO managers and the operations manager concerning VO setup and support. The web interface allows the registration of new VOs as well as the update of relevant information of approved VOs. The VO identity cards (VO ID cards) hold static information about VOs, particularly:

- Basic information (enrolment URL, status, discipline, etc.);
- Contact points (VO managers, VO user mailing list, etc.);
- The acceptable user policy of the VO;
- List of core services of the VO;
- VOMS information (groups and roles, certificate details, etc.);

- VO specific requirements of sites supporting the VO are optionally recorded (minimum CPU, RAM, etc.).

The Virtual Organization Membership Service (VOMS) URL endpoint, stored in the VO ID cards, is of the utmost importance for the framework to retrieve the number of users associated to each VO. The VOMS keeps a registry of members of one or more VOs. These members are registered in VOMS when they join the VO. The registration process requires the user to provide the 'Distinguished Name' (DN) of his/her personal X.509 certificate. These DNs are recorded by VOMS – and by the resources of the VO – to identify users when they access the sites, for example to run jobs or store files.
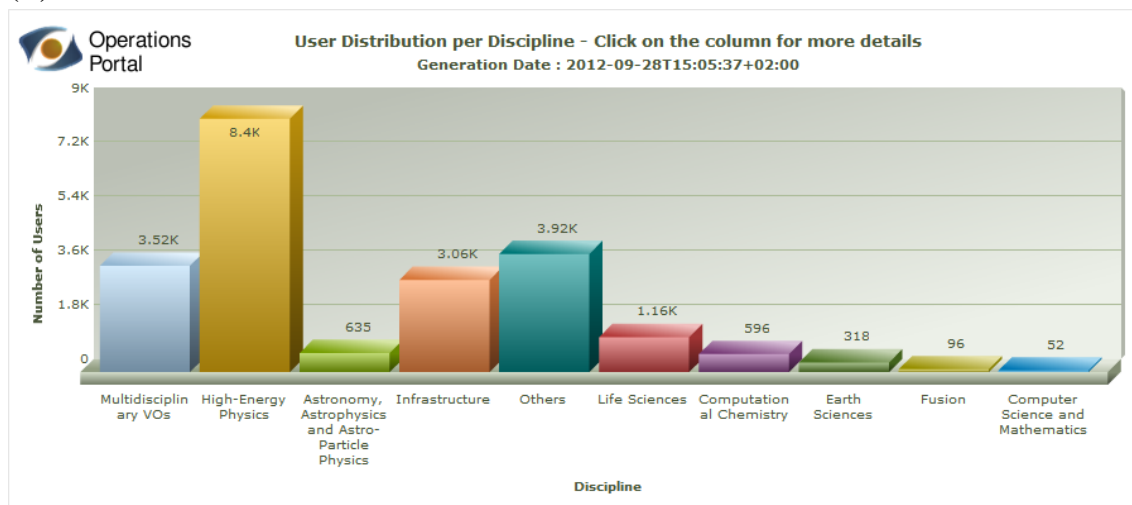
Distinguished Names are made up of multiple attributes such as the 'Common Name' (CN). The number of EGI users is the number of distinct CNs recorded in the Operations Portal data-base. For completeness, a user can have multiple X.509 certificates; at the same time, a user can be registered in multiple VO's.

3. Calculating user numbers

3.1 User statistics

Some of the most important metrics about the uptake of EGI are available through the Operations Portal, namely the spread of users across VOs, across scientific disciplines, and across certification authorities (institutes that issue X.509 certificates). The VO metrics generator tool is capable of displaying the changes of metric values on a timeline. The data is accessible either in a chart or in a table format, and can be downloaded as well for further manipulation. As an example, figure 2(A) shows the distribution of users per discipline, generated in the 'Users Summary' interface on the 28th of September 2012 [3]. The number of EGI users is steadily growing, currently surpassing the 21,000 users as shown in figure 2(B). This chart was generated from the data taken from the 'Users metrics' interface [4].
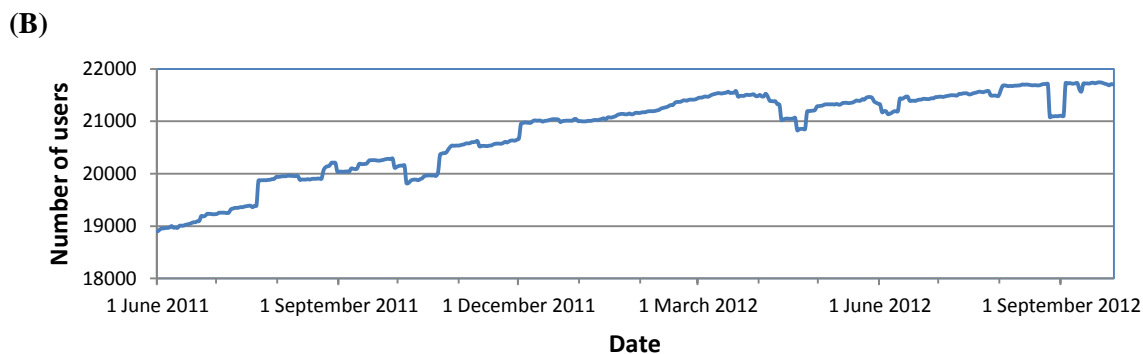
(A)

(B)

Figure 2 – **(A)** European Grid Infrastructure user distribution per discipline. Disciplines displayed from the left to right: Multidisciplinary, High-Energy Physics, Astronomy Astrophysics and Astro-Particle Physics, Infrastructure, Other, Life sciences, Computational Chemistry, Earth Sciences, Fusion, Computer Science and Mathematics. **(B)** Evolution of the total number of users.

3.2 Methodology

The workflow used by the Operations Portal to retrieve the list of users from the VOMSes is presented in figure 3, and consists of the following steps:

1. The VO managers use the Operations Portal to record VO information in the VO ID cards;
2. The information related to VOMS servers is registered in the Operations Portal VO database (CIC DB) and made available to the Lavoisier Web Service;
3. Lavoisier builds the VOMS contact URL list from the data;
4. Lavoisier contacts the VOMSes via a SOAP method and retrieves the list of registered users certificate DNs for each VO;
5. The users list is recorded in the CIC database and a cache is also kept in Lavoisier;
6. The records are aggregated and sorted to expose consistent VO user statistics on the Web Interface.

4. Identified issues

In the last couple of years the above described 'Operations Portal – VOMS' reporting workflow was the only method in use to gather statistics about EGI users. Our team has recently identified several issues that restrict the usability of the workflow. The issues are described in the next subsections.
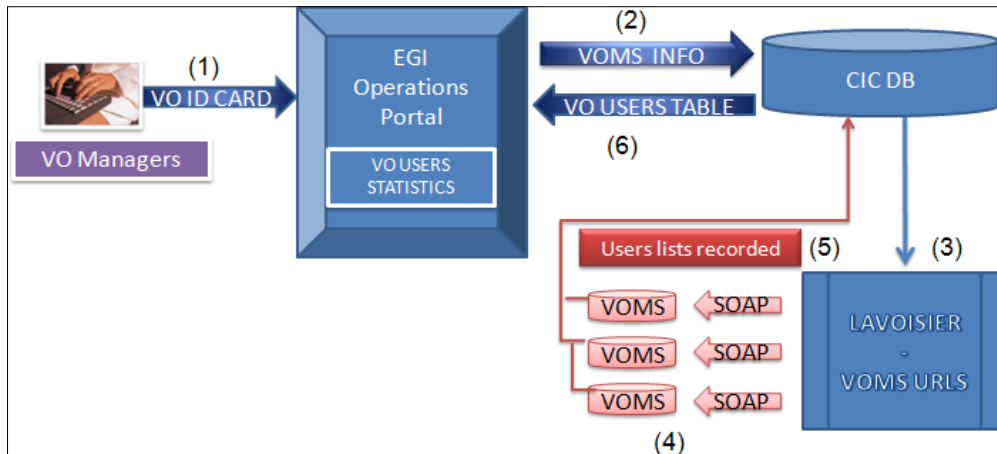
6

Figure 3 – Operations Portal methodology to retrieve number of EGI users.

4.1 Invalid VOMS information

The Operations Portal gathers user information from VOs using the VOMS servers that are registered with their URLs in the VO ID cards. The workflow does not provide the correct number of users for a VO if:

1. The VOMS registration is incorrect: the VOMS URL endpoints are missing or outdated in the VO ID card[1];
2. The VOMS servers are unavailable: VOMS servers can be down or not reachable because of erroneous installation or configuration.

4.2 Expired user registrations

The personal X.509 certificate that one needs to register in an EGI VO is usually valid for one year, with possible annual extension. When a certificate expires its registration, it is not automatically removed from the VOMS servers and thus is also accounted by the Operations Portal as an active user.

4.3 Inactive VOs

Some VOs can be inactive for a long period of time (over 6 months), which means that none of the members have performed any calculation on the grid. The VO ID cards of these VOs can be still in the registry and their users are thus included in the statistics.

---

[1] A correct VOMS URL looks like (fields in italic should be replaced accordingly):
https://*voms.address*:8443/voms/*VOname*/services/VOMSAdmin?method=listMembers

4.4 New types of access

At the time of writing almost all of the computing and storage capacity of EGI is exposed to user communities through gLite, ARC, Unicore and dCache grid middleware services. Access to these services is traditionally possible with personal X.509 certificates, obtained from certification authorities recognised by EGI. During the last decade e-infrastructure communities and their perception of certificate based access to e-infrastructures has changed significantly. Many of the existing and potential users of EGI consider the personal certificate based access as one of the main barriers of uptake. Some of these users – together with support teams from the National Grid Infrastructures (NGIs), NRENs and scientific projects – developed various solutions to simplify access to EGI resources [5]. Some of these solutions, particularly robot certificates and identity federations, completely eliminate the need for users to have X.509 personal certificates to interact with EGI. Consequently, those who use these methods when they interact with EGI do not get registered in VOMSes and remain 'invisible' to the Operations Portal.

4.4.1 Robot certificates

Robot certificates are X.509 certificates used by 'application robots' to interact with the grid infrastructure. An application robot submits jobs, reads/writes files and performs various other grid operations on behalf of its users. A secure way to store X.509 robot certificates is through security tokens, an hardware or software component which provides strong authentication operation to an authorized computer service [6]. These tokens can offer grid automated authentication - thus the name robot - to an application through a long-lived robot proxy whose private key is not protected with a passphrase. Robot certificates are often coupled to automated monitoring services, automated data transfer services, grid job submission/retrieval, amongst others.

Application robots are registered in EGI VOs with the DN of their robot certificates, just like ordinary VO users with the DN of their personal certificates. However, behind each of the robot registration there can actually be numerous users, whole research groups, whole research communities. Because these users do not register in the VOs individually, they are not counted by the Operations Portal.

4.4.2 Identity federations

In a distributed environment, many information services – such as e-mail, library databases, data repositories, portals, grid/cloud computing applications – require users to authenticate themselves. Within a single institute, an institutional identity management system can simplify this authentication for the users. Rather than having separate credentials for each system, a user can employ a single digital identity to access all resources to which he/she is entitled within the organisation. Federated identity management extends this approach beyond the institutional level, creating a trusted authority for digital identities across multiple organizations [7].

An identity federation is made of "… the agreements, standards, and technologies that make identity and entitlements portable across autonomous domains (Burton Group)". Within an identity federation, engaged institutions share identity attributes based on agreed-upon standards, facilitating authentication from other members of the federation and granting appropriate access to online resources. When a user, who is affiliated with a member of a federation, requests a protected resource from another federation member organization, he/she is prompted for identifying information including his/her 'home' organization. This request is passed on to his/her home organization, which verifies user's credentials (very often a username-password pair) and asserts to the requesting organization that the user has been authenticated. The organisations that provide 'user verification' within the federation are called 'identity providers' (IdPs in short), and the organisations that provide services that verified users can access are the 'service providers' (SPs in short). Identity providers determine individually which attributes about authenticated users will be shared, such as name, title, or role. Based on this information and their respective policies, service providers then grant or deny access to particular resources.

Identity federations can enable users to access EGI services with username-password they were given by their home institutions and what they traditionally use for accessing institutional services such as email or intranet. Several NGIs and projects already provide bridging technologies to interface identity federations with grid middleware, cloud or portal environments. Several Web portals and application frameworks were already built and are used by various communities across Europe and around the world.

In case of the federated identity access a user usually does not need to have X.509 personal certificates and register in a VO. The bridging technology can map federated identities to robot certificates (as described in section 4.4.1), or can generate other types of certificates to interact with grid infrastructure services (for example online certificate authorities). In these cases the users of the system are not visible to VOMS servers and by consequence to the Operations Portal.

5. Implemented solutions

We have implemented a series of solutions to address the above descried problems. These solutions are described in the below subsections.

5.1 Correcting VOMS information

The list of incorrect and inaccessible VOMS servers can be found from the Operation Portal logs. An email was sent to the managers of these services who were asked to correct the VOMS URL method in the VO ID card, and to check the proper behaviour of their services directly from the portal. Two sets of emails have been sent in late 2011: the first to VOs without VOMSes registered, the second to VOs with unreachable VOMSes. As a result sixteen VOMSes were corrected or added to the ID cards, and these account for 400 additional users.

Some of the VO managers have not responded to the emails. For these VOs we have extracted VOMS information from the grid information system (BDII) and updated the VO ID

cards accordingly. The outcome was the identification of 10 VOs, which after an update of their ID cards accounted for an additional 300 users to the statistics.

VOMS servers that are older than v2.0 were not compatible with the SOAP method that the Operations Portal use for user query (step 4 in figure 3). A campaign to upgrade VOMSes to v2.0 or later was run in 2011. By the time of writing all VOMSes in the Operations Portal are v2.0 or higher.

Taking into consideration the bottleneck VOMS associated errors cause while computing the number of EGI users, further development of VOMS was requested. The VOMS should be published into BDII along the number of users. In practice this means that instead of querying all VOMS servers, we will extract the information from a single entry point from the information system.

Overall, the solutions implemented regarding the VOMS information issue have improved the quality of the user metrics. Nevertheless, a continuous effort must occur to closely follow the status of all VOMS registered in the VO ID cards. From time to time some VOMS are unavailable, and the respective VO users are not statistically liable by the Operations Portal, as can be seen from the sudden drops of EGI users in figure 2(B). In such a case, we contact the VO manager to fix the problem, in order to provide proper metrics for the affected VO(s).

5.2 Recognising and excluding expired user registrations

The VOMS serves as a central repository for VO user authorization information. The membership expiration mechanism enforces that each VO member is known and approved by the VO manager in compliance to the Virtual Organisation Membership Management Policy [8]. When a user is registered in a VO, an expiration date is linked to his/her membership being the default lifetime 12 months. Membership expiration and renewal is enforced in VOMS, a feature implemented in VOMS Admin (provides tools for administering member databases for VOMS) version 2.5.3-1 in May 2010 and following releases. The access to the VOMS proxy signature interface is limited to the users that are listed as active members to the VO. Removing a user from the VO, or suspending his membership, will block his/her ability to obtain a valid proxy signature from the VOMS server [9].

EGI has an on-going middleware upgrade campaign, where all sites have to upgrade their services namely the VOMS, by the end of October 2012. The latest VOMS version 2.0.8, developed by the technology provider EMI, was certified in July 2012 by EGI as part of the Unified Middleware Distribution [10]. The VOMS admin tool includes new features such as API access to information about active, suspended and expired users. Once all sites are updated the Operations Portal will be able to trustworthily retrieve the DN's of users with a valid VO membership.

5.3 Recognising and excluding inactive VOs

The EGI Accounting Portal [11] provides statistics about resource use by the VOs. The inactive VOs, obtained from the Accounting Portal, were contacted for a confirmation of

inactivity and were eventually suspended and removed from the statistics. Approximately 10 VOs have been identified and removed, a loss of around 250 users from the statistics.

5.4 Solution to robot certificates issue

The robot certificates being used to access EGI resources is known to the Operations Portal, because the robots need to be registered in EGI VOMS servers and respective DNs have a distinguishable naming pattern easily identifiable during the SOAP queries [12]. For instance, the subject DN of a robot certificate must unambiguously identify the entity as a robot by including the string 'Robot' as part of the Common Name attribute. In July 2012, 59 robot certificates were registered in EGI VOs. Unfortunately at the time of writing there is no better way to gather statistical information about users of robot certificates than asking this information from the operators of the robots. A survey has been setup by the EGI.eu User Community Support Team in January 2012 to collect user statistics from robot owners [13]. The received answers from the January and August campaigns are visible on a public page in the EGI Wiki [14]. The response rate is unfortunately quite low (answer from ~40% of the robot owners), but based on the answers we assume that in January 2012 around 300 people had access to grid resources through robot certificates and by August this number went up to around 400-450.

6. Open issues

Despite the above described actions successfully tackled some of the problems with collecting user statistics, some of the issues remained open. This section collects these and outlines the challenges raised by the issues.

6.1 Inactive users

The list of users retrieved by the Operations Portal from the VOMS servers consists of those who have a valid personal grid certificate. Just because someone has a grid certificate and is registered in a VOMS server does not necessarily mean that this person is an active user of the grid. The EGI Accounting Portal stores information regarding the usage of resources within the EGI production infrastructure. Because of legal reasons most of the EGI sites are not allowed to expose information about individual users' resource use. Therefore the EGI Accounting Portal provides resource usage information at the VO level.

The legal issue concerning the publication of individual users' usage record is expected to remain. At the same time knowing more about the most active users, about cross-national use of the infrastructure would help the NGIs' strategic planning. The EGI-InSPIRE project therefore kicked off a series of Virtual Team projects [15] to collect more information about the NGIs' resource use, and provide a regular report giving an overview of the usage of the EGI infrastructure. The output of these projects may provide statistics that could be cross-checked with the VO membership statistics in order to identify and remove inactive users from the statistics.

6.2 Certificate-less access

Solutions, such as robot certificates or identity federations eliminate the need for end users to possess personal certificates for EGI use. There are already several high level environments (Virtual Research Environments) like portals or desktop based science gateways, which are using such solutions. The number of such environments is growing, driven by recently started projects such as SCI-BUS [16], which will establish EGI-based science gateways during 2011-2013 for around 20 research communities.

The survey based collection and Wiki based sharing of statistical information about robot certificate users (described in section 5.4) is not a long term solution. The low response rate to surveys and the difficulties in expanding it to identity federations predict an end to the method. A long term approach for assessing the number of users who access EGI resources without personal certificates would be some sort of reporting system that can be easily implemented by the Virtual Research Environments that users interact with to access EGI.

The technical-political details and feasibility of such a solution will be discussed within the recently started EGI Virtual Team project, titled 'Science Gateway Primer' [17]. One of topics to be covered by the project is to explore how gateways can be monitored by the EGI service availability framework. A set of customized probes could be developed within this framework to query the number of users served by a gateway. After the methodology and tools are established for science gateways, these can be extended to other types of environments that enable certificate-less access to EGI.

7. Conclusions

The number of EGI users being serviced is an important indicator to evaluate the uptake and success of European e-infrastructure services. Knowing the number and scientific background of EGI users is important to the NGIs, the main stakeholders of EGI. NGIs need this information for strategic planning activities and for negotiations with national funding agencies.

Traditionally it was the EGI Operations Portal that provided information about the number and scientific spread of the EGI user base. Our work identified flaws in the Operation Portal based mechanism, improved and extended it. While some of the described solutions are already in place, others are still waiting for implementation.

The work has not finished yet. Novel ways of infrastructure access bring new challenges. Emerging computing platforms complicate user calculations even more. Some of the NGIs are experimenting or are already providing services to scientific communities using alternative platforms, such as clouds or Hadoop. Some of these platforms are exposed to international use, while others for institutional or national use. The EMI and IGE projects, the two largest technology providers of EGI, will come to an end in May 2013. The list of EMI and IGE technologies that will survive these projects is still unknown. These various known and unknown factors will require the community to regularly revisit and try to answer the question 'How many users does EGI have?'.

8. Acknowledgements

References

[1] Operations Portal homepage available at: http://operations-portal.egi.eu

[2] Lavoisier Web Service homepage available at: http://grid.in2p3.fr/lavoisier

[3] Users summary interface available at: http://operations-portal.egi.eu/vo/usersSummary

[4] Users metrics interface available at: http://operations-portal.egi.eu/vo/usersMetrics

[5] G. Sipos, Authentication solutions in the European Grid Infrastructure, EGI public report, July 2012: https://documents.egi.eu/document/1178

[6] Security token definition available at: http://en.wikipedia.org/wiki/Security_token

[7] EDUCAUSE: 7 things you should know about federated identity management, Sept. 2009: http://net.educause.edu/ir/library/pdf/EST0903.pdf

[8] EGI Virtual Organisation Membership Management Policy: https://documents.egi.eu/document/79

[9] VOMS Service Reference Card: https://wiki.italiangrid.it/twiki/bin/view/VOMS/VOMSServiceReferenceCard

[10] Virtual Organization Membership Service v2.0.8: http://repository.egi.eu/2012/07/10/voms-2-0-8/

[11] EGI accounting portal available at: https://accounting.egi.eu

[12] EUGridPMA guidelines on approved robots: http://www.eugridpma.org/guidelines/robot

[13] EGI blog on users and robot certificates: http://go.egi.eu/blog.robot.users

[14] List of robots interacting with EGI resources: http://go.egi.eu/wiki.robot.users

[15] EGI Virtual Team Project on Inter-usage report: https://wiki.egi.eu/wiki/VT_Inter-Usage_Report,_part_II

[16] Z. Farkas and P. Kacsuk, P-GRADE Portal: a generic workflow system to support user communities, Future Generation Computer Systems journal, Volume: 27, Issue: 5, 2011, pp. 454-465

[17] Science Gateway Primer - EGI Virtual Team project: http://go.egi.eu/vt.sgp