Interoperable e-Infrastructure Services in Arabia

Salem Al-Agtash1,2,3, Roberto Barbera4

1Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA
2Arab States Research and Education Network GmbH, Dusseldorf, Germany
3Department of Computer Engineering, German-Jordanian University, Amman, Jordan (On Leave)
4Department of Physics and Astronomy of the University of Catania, Catania, Italy

Email: alagtash@gmail.edu

Abstract

e-Infrastructures became critical platforms that integrate computational resources, facilities and repositories globally. The coordination and harmonization of advanced e-Infrastructure project developed with partners from Europe, Latin America, Arabia, Africa, China, and India contributed to developing interoperable platforms based on identity federation and science gateway technologies. This paper presents these technologies to support key services in the development of Arabia networking and services platform for research and education. The platform provides scientists, teachers, and students with seamless access to a variety of advanced resources, services, and applications available at regional e-Infrastructures in Europe and elsewhere. Results showed that on average about 35,000 monthly jobs are running for a total of about 17,500 CPU wall-clock hours. Therefore, seamlessly integrated e-Infrastructures for regional e-Science activities are important resources that support scientists, students, and faculty with computational services and linkage to global research communities.

Keywords

e-Infrastructure, Science Gateway, Identity Federation, Computational Services

1. Introduction

Research and education are now globalized and require efficient communication
and innovative services, literally named e-Infrastructure or cyber-infrastructure. An e-Infrastructure ranges from the physical supply of research networks to providing access to data for virtual research communities. It includes organizations and services as diverse as national and international multi-purpose grids, supercomputer infrastructure, data grids and repositories, tools for visualization, simulation, data management, storage, analysis and collection, tools for support with regard to methods or analysis as well as remote access to research instruments and very large research facilities.

Today, research has increasingly become computationally and data-intensive and of inter-disciplinary nature. Therefore, new tools are needed to analyze, model and visualize diverse datasets to support a new science paradigm as well as to develop complex applications to solve pressing problems and issues. In general, these tools are made accessible on dedicated research and education networks that connect universities and research institutes at large. Providing seamless access to global distributed tools, applications, and repositories as well as sharing of grid, cloud and high-performance computing resources have become a global trend. Virtual research communities have been formed on e-infrastructures to address new scientific challenges through collaboration. Unfortunately, a variety of e-infrastructures are still isolated and context insensitive. Therefore, their interoperability to create larger computational capacities is seen critical to address the requirements of cross-continental research communities and for the development of research and education and the advancement of science and technology, worldwide, [1] [2] [3].

Interoperability is a property referring to the ability of diverse systems and organizations to work and interoperate together. It defines standards and policy guidelines for services and applications between regional e-infrastructures and integration of operations, including: storage and hosting of content delivery; services/applications for research and educational community; network communication tools and resources; and virtual learning environments. With interoperability, research communities in computationally intensive scientific areas such as genomics, climate change, and medical diagnostics can easily interact and take advantage of the distributed computing resources: scientific applications and tools, data repositories, CPUs and storage disks.

Science gateways [4], and identity federations [5], have emerged as new technologies to facilitate interoperation of e-Infrastructures and authorized access to distributed computational resources. A science gateway is a community-developed set of tools, applications, and data that is integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community. An identity federation is made of the agreements, standards, and technologies that make identity and entitlements portable across autonomous domains [6]. It has the aim of setting up and supporting a common framework for different organizations to manage accesses to on-line resources [3]. These interoperation technologies create a basis for open access, education and research and enable e-Infrastructures to connect
computing and storage resources, data and researchers from all over the world. The global knowledge bases of open access data, document and education repositories, which have been built by the CHAIN-REDS project, are meant to serve both researchers and scientists at the global level [7].

Arabia has been developing its e-Infrastructure that is made up of networks and services through close coordination with their European counterparts. Its ultimate objective is to build a more cohesive platform for robust scientific collaboration. Interoperability and coordination of these e-Infrastructures to interlink continentally and to create larger computational capacities has been identified as a critical mission in the development of research and education and for the socio-economical welfare of the region. A number of Arabian countries interconnect with regional e-Infrastructures either using the commodity Internet or have links to Europe with limited capacity [7]. This, of course, poses a limitation in capacity use and dramatically impacts on the development of research collaborations and innovation in general. Services and resources need also to be developed as well to justify the investment necessary to build up the regional e-Infrastructure. Providing Arab young scientists with viable tools and services for seamless access to first-class advanced computing resources and facilities through high-speed dedicated networks to allow them to integrate with their peers worldwide is indeed a challenging task that has not yet been given much attention.

This paper contributes to the development of tools to harmonize ubiquitous interoperation between regional e-infrastructures and help research communities to benefit from these infrastructures. The results of this paper build on the collaboration models that have been established by many projects and initiatives between the different e-infrastructures in Europe, Latin America, China, India, Africa, and Arabia in the framework of the EU FP7 funded projects [8] [9] [10]. These models provided tools for the harmonization of e-infrastructures to help researchers in the region benefit from services provided outside their local networks. It will be shown how emerging regional e-Infrastructure in Arabia has customized existing technical and organizational best practices to be able to provide services on top of the high bandwidth network and make a balance between scalability and pervasiveness. The remaining sections of this paper are organized as follows: a background of e-Infrastructure development is given in Section 2. Sections 3 and 4 provide brief technical details on the interoperation platform enabled by identity federations and science gateway. Illustration on a preliminary development is given in Section 5. Finally, conclusions are drawn in Section 6.

2. Regional E-Infrastructure

Arabia is highly regarded as a region of emerging economy and is host to many western universities and businesses, which require substantial linkage and networking with their counterparts elsewhere. It comprises 22 countries with about 350 million inhabitants and over 10 thousand institutions, networking of which
has become critical to take advantage of today’s research and education advanced resources and be able to integrate with research communities at the global level to address issues of common concern. The development of Arabia regional network and its global interoperation through variety of European funded projects, including [8] [10] [11], have tackled many of the networking challenges and issues as well as interoperability of services.

Regional e-Infrastructures exist to connect national research and education high-speed communication networks. The European GÉANT, US Internet2, Canadian CANARIE, Asia-Pacific APAN, and Latin American CLARA are examples of regional networks. So far, the coordination between the regional e-Infrastructure efforts has been restricted to basic operational, organizational and technology know-how and exchanges. The upsurge of other paradigms, such as virtualization and cloud computing, represent additional trends in the light of a global e-Infrastructure landscape. In Arabia, the evolution of e-Infrastructures, foreseen cross border connectivity, established international linkages and capacities, and created Arabian Global Exchange points in London and Fujairah are seen as major developments. These are enhanced with federated identity and science gateway platforms that are necessary to provide key services necessary to sustain Arabia e-Infrastructure. The interoperable e-Infrastructure allows young scientists to seamlessly use varieties of applications, services and repositories in a large-scale ubiquitous computing environment.

In Europe, GÉANT connects 3900 institutions in more than 40 countries and provide education, science, and research services to more than 30 million students, teachers and researchers. The Asia-Pacific Advanced Network APAN connects several countries and support interlinks and services to many scientists. In Latin America, RedCLARA has played an important role in developing the interconnection between 16 Latin American countries’ education and research networks. UbuntuNet provides coordination for network infrastructure in the Southeastern region of Africa.

3. Interoperations

Interoperations between regional e-Infrastructures allow global access to shared resources, computing services, and data repositories. CHAIN-REDS has enabled five regional operation centers worldwide to interoperate with the European Grid Infrastructure (EGI). The objective is to support the use of advanced technologies and resources and to interface with the EGI. It has provided links to data repositories, open access document repositories, and open access education repositories [9] [12] [13]. Cloud computing emerge as a comprehensive technology to provide wider access to computational infrastructures. CHAIN-REDS has also enabled cloud infrastructure and provided cloud middleware via cloud data management and open cloud computing interfaces.

A number of research communities have intensively used these services and resources and demonstrated potential use in variety of domains. Examples of
these communities are:
- The African Population and Health Research Center
- The Latin America Giant Observatory
- Genome and protein structure prediction using TreeThreader
- Bio-molecules and molecular dynamics simulation using GROMACS
- Quantum chemistry and physics of materials using ABINIT calculations based on density functional theory

Arabia is highly regarded as a region of emerging economy and is host to many western universities and businesses. The aim of the Arabia e-infrastructure is to provide faculties, researchers, and students with ubiquitous and reliable services for networking and computing as well as the open access to e-Science environments and European data resources. As the access to data-driven and computer-intensive resources and services is the basis of innovation improvement and the advancement of knowledge in the academic communities, Arabia e-infrastructure targets:
- Connecting education networks to provide scalable inter-domain services, grid and cloud infrastructures. Having such a virtual access to the services and data lacked in the Arabian institutions will enable researches to join the modern research fields like bioinformatics for example.
- Creating the needed protocols to identify priorities and to provide researchers in Arabian academic institutions with the needed processing at supercomputing facilities upon request.
- Create an open and a trusted infrastructure to hold the highly accessed scientific information to engage research data sharing and to conduct joint research activities.
- Create mechanisms, strategies, and business models for inter-commercial services to optimize and sustain related investments.
- Create a shared cross-countries multi-disciplinary innovation environment.
- Create data strategies, standards, certification schemes, and ontologies to manage data usage and to harmonize the procedures regulating the delivery of virtual research services. Accordingly, partnership of research organizations with industry can be reinforced.
- Support trans-national software implementations and trainings.
- Propose flexible trans-national business models to ensure financial sustainability.
- Develop and distribute the standards codes and know-hows among the participating research communities.
- Improve trans-national competitiveness and productivity between participating institutes and companies.
- Promote a change of culture of research communities towards open data.
- Arabia e-infrastructure is able to assemble mass of people, knowledge and investment, which contributes to national and regional economic development.
4. Identity Federation

Identity federation is based on facilitating user identities across several domains through single sign-on [14]. It is a mechanism to authenticate a user across multiple sites in a domain or across independent domains using open standards. The objective is to give a user the flexibility to federate his/her multiple accounts using one set of credentials (usually, username and password). The Liberty Identity Federation Framework is a common industry standard for the laws of federated identity and user-centric data exchanges among trusted domains. The Security Assertion Markup Language (SAML), with its implementation done by Shibboleth, is the most widely adopted standard used to exchange the authentication and authorization data [15]. New open source protocols like Higgins, policy frameworks like Shibboleth (Anon., n.d.) and decentralized frameworks like OpenID are also used as means of federating identities. Many commercial service providers offer identity management and authentication services. Facebook and Google are common social identity providers. These capabilities allow service developers to integrate their different identities into their services by enabling user identities to be sourced from several directories (e.g., LDAP) and exposes different authentication interfaces that can be embedded in external applications (e.g., OpenID).

The Research and Education Identity Federations—REFEDS—represents stakeholders from NRENs, industry, business, and research and education communities [16]. Its objective is to address the needs of research and education identity federations worldwide and promote the development of access and identity management technology, policies and processes. REFEDS is a community of practice that plays an important role in developing policies for inter-federation, privacy, assurance, and relationship with partner communities as well as for marketing and supporting emerging federations. It is actively engaged in access and identity work and supportive of standards-compliant developments to enhance international collaboration.

eduGAIN is a service available to enable seamless exchange of information related to identity, authentication and authorization between the GÉANT Partners’ identity federations [17]. The exchange is made through coordinating elements of the federations’ technical infrastructure and a policy framework. The objective is to enable Pan-European Web single sign-on to GÉANT and its GN3 partners’ services. eduGAIN interconnects identity federations around the world, simplifying access to content, services and resources for the global research and education community. It helps identity providers to offer a greater range of services to their users, delivered by multiple federations in a truly collaborative environment. It also helps service providers to offer services to users in different federations, while providing users with a wider range of services. Figure 1 shows how eduGAIN interconnects federations (eg: A, B, C) to allow cross authentication and access to services, [2] [3]. Each federation represents the interconnection between Identity providers (IdP) and services providers (SP).
Science gateways are becoming popular tools used by research communities to interact with e-Infrastructures and to access shared data, software, educational resources, and computational services [4] [18]. The tools represent community-specific set of applications and data collections that are integrated together via a web portal or a desktop application. They provide access to resources and services available at the grid infrastructure [19] [20]. The tools also allow users to analyze and share large volumes of data and to run scientific applications and simulations on grid and super computers. Science gateways can offer a variety of services including virtualization, job execution, access to data collections, and applications. Interoperability between different e-Infrastructures and integration of operations are key aspects of science gateways. A number of Arabian countries have adopted the Catania Science Gateway Framework for implementation by the CHAIN-REDs Project [9]. The project has promoted the uptake and use of Science Gateways supporting projects such as agINFRA, DCH-RP, EarthServer, and eI4Africa with 1000 people registered, 25 integrated applications, and 2500 open access document repositories.

5.1. Architecture

The generic architecture of Catania science gateways is presented in Figure 2. The architecture facilitates access from anywhere to the applications embedded and deployed in the science gateway. Applications are interfaced to the underlying Grid infrastructure through a library of software services, which is based on standards and is middleware-independent. Users in different research groups are given certain privileges to different applications and dataset modules, which are accessible through simple web-based interfaces. They are transparently authenticated and authorized on all e-Infrastructures without any additional human/machine intervention. User authentication relies on Identity Providers (IdPs) that are members of one or more Identity Federations. In this case, authentication is made through ASREN Identity Federation.
Interoperability is facilitated through science gateways. It allows diverse systems and organizations to work and interoperate together. Standards and policy guidelines are defined for services and applications between regional e-Infrastructures and integration of operations, including: Storage and hosting of content delivery; Services/applications for research and educational community; Network communication tools and resources; Grid computing and coordination; and Virtual learning environments.

With interoperability, research communities in computationally intensive areas such as genomics, climate change, and medical diagnostics can easily interact and take advantage of distributed computing resources: scientific applications and tools, data repositories, CPUs and storage disks.

5.2. High Level Interface

High-level interfaces are provided to users for quick access to distributed computing and storage resources. It provides a set of well-defined and domain specific applications for a smooth access, while preserving security imposed by the distributed e-Infrastructure and the topology of the sensible information managed. Several web and Grid technologies have been adopted and deployed for different VRCs to ensure compliance of authentication requirements. The highest component in the authorization/authentication hierarchy is integrated in the Science Gateway and supports a Single Sign On mechanism across all services a given user is entitled to use. SAML is used for the credentials communication, on the basis of Shibboleth System [21]. Identity Federations are being deployed in different regional networks to simplify access to services for users working in different locations.

When a user signs into the science gateway and is authenticated and authorized to run an application, a proxy certificate is issued to secure Grid transactions. Robot certificates are used and managed on the multi-threaded eToken server, stored in different USB eToken PRO 32/64 KB smart cards [22].

5.3. USE Cases

A number of use cases for e-Infrastructure services in different regions, covering...
different sets of user requirements, have been identified, mainly: Molecular dynamics—GROMACS, Materials science—ABINIT, Astronomy—LAGO, Population and health—APHRC, and Proteomics—TreeThreader [9]. The GROMACS software package is used for molecular dynamics simulations. This kind of studies presents a huge computational demand. 14 European, Arab and Indian Grid sites have already been enabled with two GROMACS versions. ABINIT requires massive calculations in the Density Functional Theory with direct applications in the fields of quantum chemistry and material physics. ABINIT has been installed in 6 European and Arabian sites. LAGO represents the Latin America Giant Observatory that relies on Water Cherenkov Detectors in 9 Latin American countries. APHRC represents the African Population and Health Research Centre for research in a wide range of topics related to societal health and wellbeing. This use case is mainly devoted to assigning Persistent IDentifiers (PIDs) to the wide plethora of datasets that APHRC manages and curates. This is of strategic importance as these datasets are widely used by almost every country in Africa in order to improve societal health and wellbeing. The TreeThreader is the leading method for protein structure prediction, and it is exceedingly time consuming. The code is already available to the desktop computing community, and is now made available on e-Infrastructure, with virtual machines launched from physical servers belonging to the China ROC and managed with OpenStack.

6. Preliminary Development

We give here insights on the preliminary development of the Pan-Arab e-Infrastructure through identity federation and science gateway. Deployment is taking place at sites in Jordan, Egypt, Algeria and Morocco. Figure 3 gives the

![Diagram](image_url)
architectural layout of the deployment. The science gateway server installs Lif-eray with federations made through idp.asrenorg.net and authorization through ldap.asrenorg.net. Users login to sgw.asrenorg.net, and when authenticated and authorized for use, ASREN eToken server issues the proxy certificate needed for Grid transactions. The core of the eToken server is a “lightweight” grid crypto. It holds the web services to access the smart cards and interacts with the automatic proxy renewal server. A Java multi-platform client is configured for inter-service communication via HTTPS. The eToken server is built on top of the Apache Tomcat Application Server and is configured to accept authorized requests.

6.1. The ASREN Identity Federation

The ASREN identity federation is based on the concept of setting up a common framework for the Arab education and research institutions to manage access to online resources, services and repositories. Researchers, users, scientists, and students use their own credentials at their home institutions (identity providers) for access. The federation framework facilitates authentication and authorization using the Shibboleth implementation of the SAML standard, to allow interoperation. It serves as the basis of the pan-Arab e-Infrastructure development. In this framework, we propose federations at the national level aggregating local Identity providers and service providers within universities and research institutions. The Arabian countries will establish the federations. In the meantime, when there is no national federation initiative, ASREN federation plays the role of an aggregate federation for all.

6.2. The ASREN Science Gateway

The ASREN Science Gateway builds on the EUMEDGRID Science Gateway, which has been implemented using the Catania Science Gateway Framework, to give users access to the distributed computing environment. Its basic elements are developed using standard portlets in the common Liferay portal framework supported by web 2.0 interfaces that easily integrate with many technologies. Users access the portal with specific roles and privileges and are allowed to run applications embedded in the Science Gateway. Applications are interfaced to the underlying e-infrastructure via a set of independent middleware services and are accessible through decoupled authentication and authorization processes. The Identity Federation based on SAML 2.0 provides the authentication and authorization is governed by a set of agreements. Several applications have been made available to scientists with seamless access. These include ASTRA, BES, ClustalW, CMSquares, GATE, Octave, Phylogenetics, and Sonification.

6.3. Testing Results

Testing has been demonstrated using the EMI-gLite-based e-Infrastructure layout at INFN Catania. A Java multi-platform client has been developed and configured for inter-service communication via HTTPS. In order to improve
performance, the server is built on top of the Apache Tomcat Application Server and configured to accept requests only from a set of authorized “clients” (i.e., the Science Gateways). The grid proxies generated by the server on request and are accessible by a Representational State Transfer (REST) API. The adoption of Apache Tomcat as an Application Server ensures scalability and high performance, with a cache mechanism implemented at the eToken Server.

The original pilot test-bed, built in the first phase of EUMEDGRID, has smoothly evolved into a production service counting 38 sites, for a total of about 4000 CPU cores and 600 TB of disk storage. Twenty applications from five scientific domains have been deployed on the EUMEDGRID infrastructure to be integrated in ASREN science gateway. The results of the application test show that about 35,000 jobs are running on average each month for a total of about 17,500 CPU wall clock hours.

6.4. Sustainability of E-Infrastructures in the Region

The ultimate goal is to implement, manage and extend sustainable Pan-Arab e-Infrastructures dedicated for research and education communities. The e-Infrastructure provides vital resources for the deployment of services that are authenticated for access to a large group of Arab scientists in more than 1000 institutions. Several efforts can be made to stimulate interest and increase usage of e-Infrastructures across the region. These include: 1) region-wide awareness campaign on computing resources, services, and applications that are available to scientific communities; 2) top-down approach to support the decision making process towards the integration in the global research and education networks for sharing experiences, learning from best practice models, and enhance collaboration with other regional communities; 3) stimulate government spending on research with funding focused on projects that are computationally intensive to address problems and issues of regional importance.

7. Conclusion

This paper presents an interoperable platform using identity federation and science gateway models for developing a pan-Arab e-Infrastructure to provide seamless access to e-Science resources, applications, and services. An architectural layout is given with details on the initial implementation setup. Twenty applications from five scientific domains have been deployed on the EUMEDGRID infrastructure and have been integrated in the ASREN Science Gateway. The results of the application test show that about 35,000 jobs are running on average each month for a total of about 17,500 CPU wall clock hours. Therefore, seamlessly integrated e-Infrastructures for e-Science activities in the region are becoming critical resources for e-Science activities. A pan-Arab regional e-Infrastructure has evolved building on the results of the framework of the EC funded CHAIN-REDS project which was concluded in 2015. A new phase of development of ASREN Science Gateway and Identity Federation is planned so that it
scales to support scientists, students, and faculty across the region with resources and services.

Acknowledgements

The authors would like to acknowledge the financial support of the European Commission in context of EUMEDGRID, EUMEDCONNECT, and CHAIN-REDS projects. Special thanks are to Yousef Torman, Ola Samara, Ashraf Alhuseini, Ramez Qunaibi, Mohamad Alshami, and Dr. Ahmad Bargash for their valuable contributions.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


