D4Science: an e-Infrastructure for Supporting Virtual Research Environments

Leonardo Candela, Donatella Castelli, Pasquale Pagano, and . . .

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1 – 56124, Pisa – Italy
{candela, castelli, pagano}@isti.cnr.it

Abstract. e-Infrastructure is the term coined for innovative research environments that provide modern scientists with seamless access to shared, distributed and heterogeneous resources. *Virtual Research Environments* (VREs) are applications whose constituents are dynamically borrowed from the e-Infrastructure, bound (and deployed) instantly, just at the time and for the period they are needed. In this paper we describe D4Science, an e-Infrastructure supporting the design and deployment of VREs.

1 Introduction

Today research activities require collaborations among parties that are widely dispersed and autonomous. Collaborations are often cross-discipline and require innovative research environments that make available data, processing and interaction intensive workflows to produce new knowledge able to stimulate further research.

To support such a demanding scenario a very promising approach is based on *e-Infrastructures*. By definition, an e-Infrastructure is a framework enabling secure, cost-effective and on-demand *resource sharing* [1] across organisation boundaries. A resource is here intended as a generic entity, physical (e.g. storage and computing resources) or digital (e.g. software, processes, data), that can be shared and interact with other resources to synergistically provide some functions serving its clients, either human or inanimate. Thus, an e-Infrastructure poses as a "*mediator*" in a market of resources having the role to accommodate the needs of resource providers and consumers. The infrastructure layer gives support to: (*i*) resource providers, in "selling" their resources through it; (*ii*) resource consumers, in "buying" and orchestrating such resources to build their applications. Further, it provides organizations with logistic and technical aids for application building, maintenance, and monitoring. A well-known instance of such an e-Infrastructure is represented by the Grid [2], where a service-based paradigm is adopted to share and reuse low-level physical resources. Application-specific e-Infrastructures are in their turn inspired by the generic e-Infrastructure framework and bring this vision into specific application domains by enriching the infrastructural *resource model* with specific *service* resources, i.e. software units that deliver functionality or content by exploiting available physical resources.

This potentially not-limited market of resources allows a new development paradigm based on the notion of *Virtual Research Environment*, i.e. integrated environment providing seamless access to the needed resources as well as facilities for communication, collaboration and any kind of interaction among scientists and researchers. This is built by aggregating the needed constituents after hiring them through the e-Infrastructure. In this development paradigm, the resulting research environments are considered as organised '*views*' built atop the pool of available assets, ranging from computers and servers to collections and services.

This paper presents the realisation of these two very challenging approaches in the context of the D4Science EU project.

2 D4Science Overview

D4Science[1] (DIstributed colLaboratories Infrastructure on Grid ENabled Technology 4 Science - Jan 2008-Dec 2009) is a project co-funded by European Commissions Seventh Framework Programme for Research and Technological Development involving 11 participating organizations. It aims at to continue the path that GÉANT[2], EGEE[3] and DILIGENT [3] projects have initiated towards establishing networked, grid-based, and data-centric e-Infrastructures that accelerate multidisciplinary research by overcoming barriers related to heterogeneity, sustainability and scalability.

In particular, D4Science is currently operating an infrastructure consisting of heterogeneous resources ranging from *hardware resources*, i.e. machines acting as computing and storage resources providers (in part borrowed from the EGEE infrastructure) or hosting environment supporting dynamic software deployment, to *software resources*, i.e. software packages implementing specific functions, *services*, i.e. running instances of software resources providing functions, and *data resources*, i.e. collection of compound information objects representing various kind of information.

This infrastructure is currently supporting the operation of two very large and challenging scientific communities: Environmental Monitoring and Fisheries and Aquaculture Resources Management. These scientific communities are served through three virtual organizations (VOs): Environmental Monitoring VO, Fishery Country Profiles Production System VO and Integrated Capture Information System VO. These VOs are dynamic group of individuals and/or institutions defined around a set of sharing rules in which resource providers and consumers specify clearly what is shared, who is allowed to share, and the conditions under which sharing occurs to serve the needs of a specific community. These VOs consists of various resources including collection of Earth images, satellite products, species distribution maps, reports, statistical data, and tools for processing and analyzing them.

The development and operation of the D4Science infrastructure is supported by the gCube software system [4]. gCube is a distributed system for the operation of large-scale scientific infrastructures. It has been designed from the ground up to support the full life-cycle of modern scientific enquiry, with particular emphasis on application-level requirements of information and knowledge management. To this end, it interfaces pan-European Grid middleware (gLite [4]) for shared access to high-end computational and storage resources, but complements it with a rich array of services that collate, describe, annotate, merge, transform, index, search, and present information for a variety of multidisciplinary and international communities. Services, information, and ma-

[1] `http://www.d4science.eu`

[2] `http://www.geant.net`

[3] `http://public.eu-egee.org/`

[4] `http://glite.web.cern.ch/glite/`

chines are infrastructural resources that communities select, share, and consume in the scope of collaborative Virtual Research Environments.

3 Building and Operating Virtual Research Environments

The D4Science e-Infrastructure supports the creation and management of VREs by offering mechanisms for the VRE definition, deployment and operation [5].

The definition process is organized in steps each allowing the designer to characterize different aspects of the expected VRE. These steps enable to collect the semantic information that the D4Science enabling system needs to automatically deploy the resources that are needed to operate the VRE. The identification of these steps and the dependences between them have been strongly influenced by the digital library model presented in the DELOS Digital Library Reference Model [6]. In particular, these steps aim at capturing VRE constituent elements belonging the Content, Functionality, Users, and Architecture dimensions.

Once the specification is completed, the VRE generation logic implemented by the D4Science infrastructure analyses it and derives an optimal deployment plan aiming at maximizing existing resources usage and eventually including dynamic resource generation. The infrastructure guarantees an optimal consumption of the available resources by selecting the minimal amount of them sufficient to meet its established performance and robustness criteria.

By using these mechanisms until now four VREs have been created serving very different application domains:

Fishery Country Profiles Production System (FCPPS) supports scientists in the generation of fisheries and aquaculture reports. The production of country profiles requires complex aggregation and editing of continuously evolving multi-lingual data from a large number of heterogeneous data sources. Availability of the FCPPS VRE permits to the scientists producing them to update and web-publish these vital reports as frequently as the community requires, while also having access to additional resources when needed.

Integrated Capture Information System (ICIS) supports scientists in integrating regional and global capture and distribution information of aquatic species, from a number of Regional Fishery Management Organisations and international organisations (FAO, WorldFish Center) into a common system. The VRE provides not only access to the necessary data but also a number of services for, providing a harmonised view of catch statistics and allowing the community to overlay according to pre-defined reallocation rules.

Global Ocean Chlorophyll Monitoring (GCM) offers to scientists an environment that integrates satellite data of microscopic marine plants and sea surface temperature. This environment support research on biodiversity, by facilitating process like the measuring the distribution, monitoring and modelling of phytoplankton (microscopic marine plants), the provision of forecasts of sea state and currents, the monitoring of algal blooms and marine pollution and the measuring of changes in the ocean productivity.

Global Land Vegetation Monitoring (GVM) provides a virtual environment that integrates satellite images of vegetative land cover. It facilitates specific research on how climate changes and land cover influence environmental resources. By having access to the data and tool of this VRE scientists can determine important measures like the total green leaf area for a given ground area, how much water will be stored and released by an ecosystem, how much leaf litter it will generate, and how much photosynthesis is going on.

As it should emerge from the brief description above, each of these four VREs offers an innovative collaboration environment. In this environment scientists addressing a specific problems can access a number of geographically disperse cross-domain resources of different nature and operate with them as if these resources were belonging to their own organization (although in the limits imposed by the resources regulating policies).

4 Concluding Remarks

e-Infrastructures that provide application services for a range of user communities cannot ignore the diversity and specificity of their requirements. In this paper, we have argued that such requirements can be conveniently met by implementing a development approach based on Virtual Research Environments and briefly presented how this approach has been put in place in the context of the D4Science project.

Acknowledgments This work is partially funded by the INFRA-2007-1.2.2 - Deployment of eInfrastructures for scientific communities - Research Infrastructures - Seventh Framework Programme of the European Commission in the context of the D4Science project (Grant Agreement no. 212488).

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organization. The International Journal of High Performance Computing Applications **15** (2001) 200–222
2. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, Global Grid Forum (2002)
3. Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., Langguth, C., Manzi, A., Pagano, P., Schuldt, H., Simi, M., Springmann, M., Voicu, L.: DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. International Journal on Digital Libraries **7** (2007) 59–80
4. Pagano, P., Simeoni, F., Simi, M., Candela, L.: Taming development complexity in service-oriented e-infrastructures: the gcore application framework and distribution for gcube. Zero-In e-Infrastructure News Magazine **1** (2009) 19 – 21
5. Assante, M., Candela, L., Castelli, D., Frosini, L., Lelii, L., Manghi, P., Manzi, A., Pagano, P., Simi, M.: An Extensible Virtual Digital Libraries Generator. In Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J., eds.: 12th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2008, Aarhus, Denmark, September 14-19. Volume 5173 of Lecture Notes in Computer Science., Springer (2008) 122–134
6. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries. DELOS: a Network of Excellence on Digital Libraries (2008) ISSN 1818-8044 ISBN 2-912335-37-X.