

## Open Source Environments for Collaborative Experiments in e-Science

Andrea Bosin<sup>1</sup>, Nicoletta Dessi<sup>1</sup>, Maria Grazia Fugini<sup>2</sup>, Diego Liberati<sup>3</sup>,  
Barbara Pes<sup>1</sup>

1 Università degli Studi di Cagliari, Dipartimento di Matematica e  
Informatica,

Via Ospedale 72, 09124 Cagliari  
andrea.bosin@dsf.unica.it,  
{dessi,pes}@unica.it

2 Politecnico di Milano, Dipartimento di Elettronica e Informazione,  
Piazza da Vinci 32, I-20133 MILANO  
fugini@elet.polimi.it

<sup>3</sup> IEIT CNR c/o Politecnico di Milano, Piazza da Vinci 32, I-20133  
MILANO  
liberati@elet.polimi.it

Open Source Software (OSS) for e-Science should make reference to the paradigm of a distributed surrounding over a multi system mix of Web Services and Grid technologies, allowing data exchanging through services, according to standards in the area of the Grid and of Service Oriented Computing (SOC). In fact, biologists, medical doctors, and scientists are often involved in time consuming experiments and are aware of the degree of difficulty in validating or rejecting a given hypothesis by lab experiments. The benefits of OSS for e-Science consider that as many operating nodes as possible can work cooperatively sharing data, resources, and software, thus avoiding the bottleneck of licenses for distributed use of tools needed to perform *cooperative scientific experiments*. In particular, this chapter presents an architecture based on nodes equipped with a Grid and with Web Services in order to access OSS, showing how scientific experiments can be enacted through the use of a cooperation among OSS sites. Such a choice, besides reducing the cost of the experiments, would support distributed introduction of OSS among other actors of the dynamical networks, thus supporting the awareness about OSS and their diffusion. An OSS environment for cooperative scientific experiments (*e-experiments*) can effectively support the distributed execution of different classes of experiments, from visualization to model identification through clustering and rules generation, in various application fields, such as bioinformatics, neuro-informatics, tele-monitoring, or drug discovery. By applying Web Services and Grid computing, an experiment or a simulation can be executed in a cooperative way on various computation nodes of a network equipped with OSS, allowing data exchange among researchers. Our environment formalizes experiments as *cooperative services* on various computational nodes of a grid network. Basic elements are models, languages, and support tools creating a virtual network of organizational responsibility of the global experiments, according to rules under which each node can execute local services to be accessed by other nodes in order to achieve the whole experiment's results.

The OSS environment supports the simulation/execution of different classes of experiments in bioinformatics (drug discovering, micro-array data analysis,

molecular docking). Each experiment ranges from visualization (browsing and search interfaces), to model identification through clustering and rules generation, and requires tools for correct design and set up of the experiment workflow and for information retrieval (e.g., for searching similar protocols, or descriptive datasheets for chemical reactors). Cooperating scientists who perform joint experiments may require specialized tools (e.g., data mining, or database tools) or computational power (e.g., long calculi for protein analysis based on their forms, or for discarding the irrelevant experiments in drug study) available only at specific nodes. The visualization part is given special attention, considering friendly interfaces and graphical simulations enabling an improved comprehension of currently textual explanations. Also privacy and data security are a major concern in our environment, considering both methods to select trusted nodes within the cooperation network, and/or to obscure or encrypt the transmitted and stored data, to preserve their sensitivity, according to user-formulated security requirements.

The environment supports experiments defined as numerical evaluations carried out on selected data sets according to available methodological approaches. The experiment execution platform is composed of: 1) a distributed workflow; 2) the involved nodes and their relative roles in the experiment; 3) the set of involved resources (data areas, data repositories and e-services). Experiments have some portions, both of processes and of data or knowledge, that can be shared in a collaborative environment. One of the reasons for executing an experiment in a distributed way might be that one organization would need to process data under a specific costly product available on a node because of its lack of skill for developing or using open source equivalent; rather than acquiring the product (e.g. SAS, ORACLE, or MATLAB), the organization might invoke a remote service as OSS available on the remote node. Another reason is that some cooperating organizations might want to inspect data dispersed on their databases, with no changes to their local computational environment. Although heterogeneous services can provide similar capabilities, the researcher is in charge of choosing the most suitable methods to accomplish each task, that is, he is in charge of designing the workflow of the scientific experiment. The run time OSS platform performs service discovery, meeting functional and non functional (e.g., price, security, quality) parameters. If the researcher wants to rerun an experiment, the workflow must take into account the changes in the choice of methods as well as in the availability of services and tools. The researcher interacts and chooses services, workflows, and data in an experimental scenario whose cooperative framework has been defined to extend the integration of scientific experiments to a level of scenario-based interaction. This scenario is profitable for many reasons, like exchanging scientific data and processing tools which results in a reduced number of software acquisitions, load balancing work between specialized researchers, and so on.