Scientific Information Management in Collaborative Experimentation Environments
Kaletas, E.C.

Citation for published version (APA):
Chapter 1

Introduction

Recent advances in technology, leading to the development of sophisticated devices, have opened the doors for scientists to perform larger and more complex experiments that were not possible before. However, new ways of experimentation have posed many requirements and challenges to both scientists and developers of supporting tools and infrastructures. New solution methods need to be designed to fulfill the needs, solve the emerging problems, and generate new results.

Tackling these challenges requires long-term multi-disciplinary efforts between scientists from both experimental science domains and computer science. In this dissertation, scientific problem solving from the computer science point of view is addressed through the virtual laboratory paradigm. Furthermore, the main body of the research work presented in this thesis focuses on the information management challenges within the scope of a multi-disciplinary virtual laboratory environment. This introductory chapter describes the emerging e-science paradigm, defines and characterizes the e-science domains and applications. It provides some examples of emerging scientific experiments to motivate and illustrate the difficulties that scientists face during their experimentations and introduces the virtual laboratory solution. Furthermore, the main objectives of the thesis are presented and the research approach to achieve these objectives is outlined. Also, the research and development projects that provided the base and context for the research work and contributed to the results obtained are described. This chapter concludes with an overview of the structure of this thesis.

Please notice that the work presented in this thesis is inter-disciplinary and covers diverse areas, including information modelling and data management techniques, distributed databases, and support environments (e.g. virtual laboratory) for experimental scientific domains (e.g. life sciences). However, the main work and focus of the thesis falls within the area of information modelling and information management, and thus covering the description of basic concepts and terminology, as well as the basic methodologies applied in the area of databases and information management is outside the scope of this thesis. The list of acronyms/abbreviations and their descriptions as well as the glossary of the key terminology, however, are included to the thesis to facilitate the reader.
It must be noted that the research approach, presented in Section 1.4, addresses the applied scientific methodology for management of information in this thesis, including the validation/verification approach applied to the scientific work of the thesis. As detailed in Section 1.4, the scientific methodology for research in the area of information modelling and information management comprises a number of steps and phases, including the field study, survey of related research, generalization of the findings to design a solution (including the concepts, models, architecture and implementation), verification/validation of the proposed solution, and the feedback to the design stage for subsequent adjustment and further improvements. Activities in different phases of this methodology are based on both theoretical and empirical work. The initial ideas are based on the literature and experiences gained from several projects. In specific, the experience gained during this research work has played an important role in the development of the models, design, and implementation of the presented information management platform. More experience was also gained through the involvement in several other research projects, which are introduced in Section 1.5 of the thesis and summarized in Appendix B. Each of these projects focuses on different aspects of information management and database software development, by examining the application and its specific data/information, identification of user requirements, modelling of the concepts/entities/functionality, design of the architecture, and choosing the implementation approach and technologies.

Validating research in the area of information modelling and information management is quite a challenge, as there are no standard means for performing it, except for empirical verification. This type of verification relies on case studies for validating the initial ideas with respect to the problem in hand, as well as for identifying the strength and weaknesses of the approach. The weaknesses identified during this process are used for further improvement and appropriate adjustment of the solution, which in turn results in further improvement of the models and the databases. Therefore, the approach to modelling and the developed information management platform of this thesis were also validated using case studies, where two scientific experimentation domains (DNA microarray and material analysis for complex surfaces) were chosen as the application case studies. One of these studies, the DNA microarray, is presented in details in Chapter 6. These case studies have provided ample grounds for the identification of basic requirements during the field study, for evaluation and validation framework that embodies both the information management approach and the models, as well as for the adjustment and fine tuning the solutions produced by the thesis.

1.1 E-Science Paradigm

In recent years, technological advances and systematic growth of research efforts in experimental science domains, such as life sciences, have achieved breakthroughs that go beyond the imagination of mankind even a decade ago. Advances in the laboratory techniques and the introduction of highly automated laboratory instruments have caused both the automation of many steps in scientific experiments and the generation
of very large amounts of data. Especially the large amounts of data have become a common characteristic of today's experiments, sometimes even leading to a change in the way that scientists perform their scientific research.

Provision of the advanced one-of-a-kind devices, the required powerful processing facilities, and the transportation, storage, management and analysis of increasing amounts of data are some examples of what is required for emerging scientific experimentations. However, these requirements can no longer be supported by a single organization. Several organizations all over the world must access to and use resources from each other. The main resources in such a networked environment comprise computing, storage and networking facilities; instruments and other experiment specific devices; raw data, produced information and published results; and analysis software and visualization equipment. The e-science paradigm provides a new way of performing scientific research using advanced computing, information and communication technologies. It is driven by large amounts of data originated from a variety of sources.

In the remaining of this section, definition and characteristics of e-science domains as well as different experiments and applications from these domains are presented. Furthermore, brief descriptions of these applications are provided as examples, and the main difficulties that scientists face when carrying out these experiments are outlined as motivation.

1.1.1 Definition and Characterization of E-Science Domains and Applications

E-science domains include the experimental science domains in which advanced computing, information, and communication technologies are extensively used for experimentations. E-science domains can be characterized as follows:

- Large amounts of data are generated by either simulations or ‘networked’ instruments (i.e. instruments that are connected to storage and computing facilities through computer networks).
- Many steps in experiments are automated; for instance, re-plating biological sample by using a pipetting robot.
- Information and Communication Technologies (ICT) are extensively used throughout the entire experiment life-cycle, from experiment design and execution to analysis and interpretation of results.

In this direction, life sciences, physics, simulation and modelling, medical sciences, and astronomy can be named among the most active e-science domains. The following outlines some examples of the emerging experiments and applications from these domains that create new challenges for the ICT. These specific examples also represent the main “problem areas” studied for the research performed in this dissertation. These experiments and applications are described in the next subsection.

- Microarray experiments and confocal microscopy experiments (life sciences)
• Material analysis experiments for complex surfaces and Large Hadron Collider (LHC) experiments (physics)

• Traffic simulation applications and fluid dynamics applications, and interactive simulated vascular reconstruction application (simulation and modelling)

Please note here that the term *emerging experiments and applications* has different meanings in different e-science domains. For instance, the emerging life sciences experiments represent a relatively new research area when compared to the experiments in physics area. However, advances in the technology has also resulted new types of experimentations in physics which were not possible before, as in the case of LHC experiments in the field of high energy physics. The common denominator for all these experiments and applications is the extensive utilization of the ICT for their support.

Four specific experiments in different domains are used as the main test cases for the research in this thesis, including microarray, confocal microscopy, material analysis for complex surfaces, and traffic simulation experiments. However, this research was also complemented with an extensive study of several other experiments and applications from the literature, which allowed the identification of the most common characteristics of emerging scientific experiments. Following are the common characteristics applicable to most, if not all, types of experiments from different e-science domains:

**Diversity.** Availability of new instruments, new techniques and new information allows scientists to gain deeper insights on problems, hence to focus on very detailed and specific questions. As a result, increasingly diverse experiments are being performed even in a single domain.

**Complexity.** In parallel to the increasing complexity of utilized instruments and infrastructures, experimental procedures are becoming longer, more complex, and more computation intensive during which various protocols and techniques are applied at different steps.

**Large data sizes.** New generation high-throughput experimental technologies, supporting technologies for the automation of many tasks during experimentation, and improved laboratory techniques have resulted in an explosion in the amount of data/information generated by experiments.

**Heterogeneity.** By nature, experiments themselves as well as the data/information they generate are heterogeneous. This heterogeneity further increases with ad-hoc experimentations driven by ‘what if’ questions, and with new types of information generated by new types of instruments and new types of experiments.

**Collaboration needs.** The need for unique and/or expensive instruments and high-capacity/high-performance infrastructure necessitates sharing of resources, experience and costs. The resulting collaborations generally include several autonomous organizations. Thus, on one hand the issue of access rights and proprietary resources and on the other hand the distribution of tasks towards joint
problem solving among different organizations are significant for proper handling.

1.1.2 Target Problem Domains and Applications – Motivation

In this subsection, brief descriptions of the above-mentioned emerging scientific experiments and applications are provided. In addition, the major difficulties that scientists face during experimentation are enumerated for each of these experiments and applications.

Microarray experiments. Among many others in the life sciences domain, microarray experiments [1, 2, 3, 4] can be named as one of the most recent growing types of experiments. Microarray experiments allow genome-wide monitoring of changes in gene expression levels to answer the question of, what genes are expressed in a particular cell type of an organism, at a particular time and under particular conditions. Changes in gene expressions occur often in response to some stimuli, for instance, the expression of diseased genes in response to a certain drug. Microarray technology exploits the preferential binding of complementary single-stranded nucleic acid sequences. A microarray is typically a glass slide, onto which DNA molecules are attached at fixed locations called spots. With the current technology of DNA microarrays, it is possible to spot up to 40,000 sequences on a single array. For gene expression studies, each of these sequences ideally identifies one gene in the genome. The microarray technology and microarray experiments are described in Chapter 6.

Main difficulties in performing microarray experiments:

⇒ Long and complex experimental procedures (e.g. a typical microarray experiment consists of 5 phases and only the clone preparation phase of an experiment contains up to 51 steps).

⇒ Archival of experiment setup, conditions, and experiment results as well as the proper links among them.

⇒ Different levels of quality requirements for different stages of experimentation.

⇒ Accessing extremely diverse set of biological data resources necessary for comparison and interpretation of experiment results. For instance, there are at present 281 biological database sources all around the world that are listed in [5], classified in 18 different categories.

⇒ Security of proprietary information, for instance to support the privacy of experiment setup and results for drug discovery studies in a pharmaceutical company.

Confocal microscopy experiments. Confocal microscopy is another emerging experiment from the life sciences domain. Confocal microscopy [6, 7, 8] is valuable for obtaining high resolution images and 3D reconstructions of a variety of biological
specimen. Confocal microscope scans the sample in slices (3D scanning), often during a time period, generating 4D images of the specimen. However, depending on a number of other parameters, such as the life time of the specimen and spectra, the image can be N-dimensional. These images are used to extract the interesting features of the object under study. For instance, in case of a time series experiment, changes in the size of a cellular organelle (e.g. mitochondria) in response to inadequate nutrition can be observed. The observations from the image stack can be combined with and correlated to the results obtained from other complementary experiments. For instance, results of a microarray experiment can be used to see the changes in the expression levels of all genes in the object of interest during the microscopy experiment period, allowing the scientist to correlate the physical changes to the molecular changes.

Main difficulties in performing confocal microscopy experiments:

⇒ Storage and management of large number of high-resolution images.

⇒ Cost and availability of software tools and hardware infrastructure required for the analysis of images.

⇒ Integrated access to heterogenous information for correlating microscopy observations to results from other types of experiments.

Material analysis experiments. Material Analysis for Complex Surfaces (MACS) experiment is an example emerging physics experiment [9, 10]. These experiments try to identify and determine the elements that compose complex surfaces regardless of the nature of the sample, and involve large and complex instruments such as the Fourier Transform Infra-Red (FTIR) imaging spectrometer, the nuclear microprobe (micro-beam), and the mass spectrometer. During the experiment phase, a sample is extracted from the source object under study, and scanned using the above-mentioned devices [9]. The large amounts of data produced by these devices make the analysis phase longer and more effort consuming than the experiment phase itself. In addition, a set of analysis tools needs to be integrated into the application to facilitate the work of scientists, such as tools for correlation analysis and multivariate data analysis (e.g. principle component analysis). Material analysis and confocal microscopy experiments are similar in the sense that both experiments focus on generating and analyzing multi-dimensional images of a sample.

Main difficulties in performing material analysis experiments:

⇒ Remote control and operation of complex laboratory instruments.

⇒ Storage and management of large number of high-resolution, large-size images (up to 100 MB per image, 20 images per day).

⇒ Combining a number of analysis tools and correlating wide variety of analysis results from a series of (possibly different types of) experiments.
LHC – Large Hadron Collider experiments. Among other emerging scientific experiments in the physics domain, one can mention the LHC [11] experiments, which bring protons and ions into head-on collisions at higher energies than ever achieved before, thus allowing scientists to penetrate still further into the structure of matter and recreate the conditions prevailing in the early universe, just after the “Big Bang”. Within the context of LHC studies, the ALICE Collaboration is building a dedicated heavy-ion detector to exploit the unique physics potential of nucleus-nucleus interactions at LHC energies. The aim of ALICE – A Large Ion Collider Experiment [12, 13] is to study the physics of strongly interacting matter at extreme energy densities, where the formation of a new phase of matter is expected. Another experiment that has been approved for the LHC is the ATLAS – A Toroidal LHC Apparatus, targeting to perform measurements that will lead to an understanding of the mechanism of electroweak symmetry breaking [14, 15, 16]. Other emerging experiments from physics and astrophysics are the Laser Interferometer Gravitational-Wave Observatory (LIGO) experiment [17] and the Astrophysics Simulation Collaboratory (ASC) [18].

Main difficulties in performing LHC experiments:

- Very high requirements for storage capacity (e.g. 100 TB of data generated by ALICE per year), network bandwidth (e.g. 2.5 GB/s event builder bandwidth and 1.25 GB/s data storage bandwidth requirements of ALICE), and computing power (e.g. 220 Gflops estimated computing power for the analysis of ALICE experimental data) [13].

- Large collaborative initiatives for sharing and reducing the costs of the required unique and expensive instruments, and high-capacity and high-performance infrastructure (e.g. the ALICE Collaboration includes 1223 collaborators from 85 different institutes of 27 countries [12], while the ATLAS project is an international collaboration involving participants from 34 countries [14]).

Simulation experiments and applications. An example engineering simulation application is the traffic simulation. Simulation of traffic flow for transportation planning [19], modelling and simulation of automatic debiting systems for Electronic Fee Collection (EFC) on highways [20], simulation of the reliability of communication links used in an EFC system [21] and calibration of the traffic generator used in this simulation [22] can be named as more detailed examples of traffic simulation.

In fact, simulation is a part of every area covered by e-science. As an example of other simulation applications, simulating fluid dynamics [23] from the domain of complex systems simulation and modelling can be mentioned.

Interactive simulated vascular reconstruction [24] is an example of emerging medical applications. The purpose of vascular reconstruction is to redirect and augment blood flow or perhaps repair a weakened vessel through a surgical procedure, where a pre-operative surgical planning allows an evaluation of different procedures a priori, under various physiological states. The aim of this application is to provide a surgeon
with an environment in which s/he can explore the effect of a number of different vascular reconstruction procedures before it is put to practice.

**Main difficulties for simulation experiments and applications:**

⇒ Size of the simulation can become very large, depending on the number of simulation parameters and required simulation runs.

⇒ Large computing power is required to run such large simulations.

⇒ Access to several databases for querying the simulation data sets may be needed (e.g. as in the case of studies for calibrating the traffic generator).

⇒ Superior visualization devices are needed (e.g. as in the case of interactive simulated vascular reconstruction where the CAVE – Cave Automatic Virtual Environment is used).

⇒ Security of private information must be provided (e.g. privacy of personal medical data in interactive simulated vascular reconstruction applications).

The following provides an overview of some of the major difficulties and pressing challenges that scientists face and need to overcome when performing their e-science experiments and applications (presented in no specific order):

- Using a variety of complex resources (e.g. laboratory instruments, computing facilities, or analysis software); sharing resources with collaborating partners; designing experiments that are capable of efficiently utilizing these (possibly distributed) resources.

- Organizing the overall working environment to avoid chaos during complex experimentations, for instance keeping track of what information is stored where; customizing the working environment to personal preference.

- Storing all information related to an experiment, including descriptions of experiment steps, interpretations of experiment results, and gained knowledge; sharing this information and knowledge with colleagues and collaborating partners; transferring the gained expertise and knowledge to inexperienced scientists.

- Ensuring the security of proprietary resources.

Clearly there is a pressing need for enhanced support environments providing adequate facilities and functionality for scientists to assist them with their complex experiments and applications. The scientific and technological challenges posed by these emerging scientific experiments and applications generally require a size and scale of effort beyond the capacity of a single physical laboratory. Human resources and expertise required for the scientific and technical problem solving may be physically distributed among two or more institutions. Furthermore, nowadays to carry out research, it is necessary and cost-effective to share access, via remote means, to
scientific instruments that are unique, scarce or otherwise difficult to access [25]. Addressing these difficulties and challenges require long-term multi-disciplinary efforts between scientists from both experimental science domains and computer science. In experimental science domains, scientists involved in these experimentations must both advance their models and mechanisms supporting the analysis of resulted information, and design innovative investigation approaches leading to discoveries [26]. In fact the complexity of the e-science experimentation paradigms necessitates that the support environment be generic and constitute a horizontal base infrastructure (serving as an operating system for this paradigm), on top of which the supporting vertical software tools can be developed, each providing a necessary service for the e-science environment. Therefore, the main challenges for horizontal infrastructure fall in the following areas:

1. data/information modelling and management;
2. experiment definition/execution infrastructure and its supporting tools;
3. high-performance computing and networking architecture and facilities;
4. interoperation and collaboration infrastructures.

At the same time, the challenges for vertical support tool developments fall in design and development of user friendly tools that best support scientists with their specific requirements. While creation of such a support environment is not straightforward within each area of science, a further challenge would be to provide a multi-disciplinary support environment for e-science. In this thesis, first a multi-disciplinary virtual laboratory environment is proposed and presented as one solution to address the computer science related challenges, and then the information management support is described and focused. The proposed virtual laboratory aims at the provision of a proper hardware and software assisting infrastructure, which supports systematic definition and carrying out of experiments. Furthermore, it allows management and publishing of information, provision of tools such as simulation and visualization tools, accessing to complex instruments, and sharing of expensive resources among various individuals or research centers. As such, the virtual laboratory helps scientists to reduce their experimentation costs and enables them to focus only on their experiments by providing the necessary supporting infrastructure. In the following section, a detailed description of the 'virtual laboratory' solution is provided. Further characterization of the virtual laboratory solution proposed by this thesis is presented in the next chapter.

1.2 Virtual Laboratory Solution

Several supporting environments for experimental sciences have been developed and many are under development, aiming to support scientists to overcome the obstacles that they face during experimentation. Some of these are targeting specific applications, while others are aiming at a generic environment supporting a (wider) range of
applications. Although science portals, problem solving environments and virtual laboratories are not identical, most of these supporting environments are offering similar or complementary functionalities.

This section provides general descriptions of different types of supporting environments for e-science domains (namely science portals, problem solving environments, and virtual laboratories). The remaining of this thesis, however, focuses on virtual laboratories.

### 1.2.1 Support Environments for Scientific Experimentations

A general overview of science portals, problem solving environments, and virtual laboratories are presented in this subsection.

#### Science Portals

*Science Portals* are emerging as convenient mechanisms for providing a single point of access with familiar and simplified interfaces to a specific set of resources that are of importance to a specific scientific community [27], typically through the use of Web browsers [28]. Resources can be, for instance, computational, storage, networking resources, electronic whiteboards for the scientific community, or a digital library.

Resources available through a science portal are usually made available to its users as services. Most science portals implement a common set of services (e.g. job submission, file transfer) [28]. These services are in general designed to provide high-level interfaces to resources, hiding the details of using the actual resource. Users can make use of these services either in a stand-alone manner using only one service at a time, or in a collective manner using a number of services simultaneously. It is also possible to combine a number of services into a single service for aggregate functionality.

Science portals only provide a uniform means for accessing resources. Users themselves are responsible for the correct and efficient usage of the available resources, either manually (in interactive or batch mode) or through an application.

The CLRC Data Portal [29], Gateway [30], UNICORE [31], Virtual Sky [32], Enter the Grid [33], the LAPK Portal [34] and the UCSD Telescience Portal [35] can be named among the examples of science portals. A classification and summary of Grid computing environments is given in [36], while the Web page for the Grid Computing Environments Research Group can be found in [37].

#### Problem Solving Environments

A *Problem Solving Environment* (PSE) is a computer system that provides all the computational facilities needed to solve a target class of problems [38]. Common characteristics shared by different implementations of PSEs are described in [39], and include among others: a target class of science or engineering design problems; natural appearance and ease of use for people using the PSEs in the target application areas, both in user interaction and ways of thinking; provision of multiple solution paths or algorithms for the entire range of applicable problems in the target areas;
and availability of parameterized algorithms to account for problem features, solution requirements, and resource availability.

Features generally provided by PSEs allow users to, among others, solve simple or complex problems, support rapid prototyping or detailed analysis by exploiting technologies such as interactive color graphics, powerful processors, and networks of specialized services [38]. Users can also be assisted through automatic and semiautomatic selection of a proper solution method from the available set of solution methods, or by providing ways to easily incorporate new solution methods [38].

The main design goals for PSEs are to increase availability of software, reliability and software lifetime while decreasing difficulty and cost/time of problem solving [40]. The study in [41] suggests that developing a robust abstract (i.e. more generic) PSE is helpful in reducing the effort required to develop and maintain robust user-specific PSEs. However, although some examples of generic PSEs do exist such as Cactus [42, 43], most of the current PSEs still target a specific class of problems. Furthermore, there are still a number of barriers and open research issues that need to be addressed to achieve more generic PSEs [44].

A PSE can make use of science portals as gateways to a number of existing resources to fulfill the solution requirements, such as computational and storage requirements.

Examples of PSEs include Cactus [42], The Astrophysics Simulation Collaboratory (ASC) [45], PELLPACK [46], Ecce [47], Sieve and Symphony [48], SCIRun, BioPSE and Uintah [49]. Also [50], [51], and [52] provide information on and examples of PSEs.

Virtual Laboratories

Another support environment for experimental sciences is the virtual laboratory. A virtual laboratory provides the necessary hardware and software infrastructure to help scientists to systematically define and carry out their experiments. Virtual laboratories are described in detail in the next subsection.

1.2.2 Virtual Laboratory

A Virtual Laboratory (VL) is sometimes also referred to as “collaboratory”. The commonly agreed definition of a virtual laboratory/collaboratory is given by William Wulf, as

“center without walls, in which the nation’s researchers can perform their research without regard to geographical location - interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries” [53].

VL provides an electronic workspace for distributed collaboration and experimentation in research or other scientific creative activities, to generate and deliver results using distributed information and communication technologies [25]. It supports an
aggregation of people who pursue a related set of research activities and share resources, where the resources including the people may be geographically distributed and associated with different institutions [54]. Therefore, virtual laboratories bring together best combination of skills, expertise, and tools to carry out the same type of research that is done in a single real laboratory [54].

VL can be used for virtual experimentation and can provide problem solving environment to scientists. It can enable a group of scientists to collaboratively work on problems under study, and assist them throughout their experimentations [26]. However, as described in [25], the supported functionality and experimentation environment provided by a virtual laboratory may differ depending on the forces driving the formation of the VL. For instance, in the case of a project driven VL, it has a well-defined goal and a fixed time scale, and it offers only the necessary functionality until the goal is achieved (e.g. an environmental research VL designing a plan for cleaning up a contaminated lake). On the other hand, in a second case several larger and more complex projects may require VL functionality for accessing large-scale facilities and for collaboration to perform multi-disciplinary activities that involve expertise from many institutions (e.g. a VL for analysis of impacts from predicted global and regional climate changes). There is also a third case where virtual laboratories can be generic and not necessarily associated with specific projects. A VL can be seen as a permanent facility providing generic functionality to its multi-disciplinary user community, including for instance remote controlling of instruments, running analysis and visualization tools, collaborating with colleagues through secure data sharing and exchange mechanisms and using video-conferencing, whiteboards, and shared applications.

At present, the concept of ‘virtual laboratory’ has been associated with several different meanings, depending on the wide variety of application interests and fields (e.g. education, games, science, engineering, manufacturing, etc.). Ongoing design and development work in the area of virtual laboratory is mostly focused on research on certain specific aspects, for instance related to overcoming the distance problem. Here, the VLs introduce mechanisms to remotely control devices, for video-conferencing, and for file sharing e.g. to enable co-working among users [55]. In the area of education, for instance, virtual laboratories can be applied to ‘virtually’ start a chemical reaction, or to see the basic mechanics rules at work while sitting in front of the computer screen [56, 57, 58]. In some other areas, virtual laboratories are used as a simulator to study dangerous situations, as in the field of aerospace [59].

In e-science, although in most cases virtual laboratory is considered as a ‘virtual’ environment where laboratory experiments can be performed in-silico and addressing a specific set of problems, there are recently some VL projects that target the provision of generic solutions and offer a wide variety of functionality, and that are applicable to a wider set of problems [60, 54, 61, 62, 63]. For instance, under the Distributed Collaboratory Experiment Environments (DCEE) program of the DOE (Department of Energy) [64], two inter-related sets of projects are supported, some focused on the testbeds and some being the technology projects. The main focus of the program can be summarized as enabling remote operations of physical apparatus and providing a history of activities through a laboratory notebook. The projects in this program include the LabSpace - A National Electronic Laboratory Infrastructure, Distributed
Computing Testbed for a Remote Experimental Environment, and the Collaboratory Development in the Environmental and Molecular Sciences. Detailed descriptions and references to each of these DOE projects can be found in [64]. Another such virtual laboratory that offers generic solutions to scientists is the Grid-based Virtual Laboratory Amsterdam (VLAM-G) developed at the University of Amsterdam, which is the subject of this thesis. VLAM-G is addressed in more details in Subsection 1.5.1.

Other examples of virtual laboratories include The Virtual Laboratory [65], Tele-Actor [66], and Virtual Lab [67].

Figure 1.1 shows how the support environments are related to each other. Please notice that a wide variety of VL configurations/architectures can result from Figure 1.1. For instance, access to resources can be handled in various ways: The VL can access the resources either through a set of science portals, or it can directly access the resources. Moreover, a middleware technology for resource management such as Grid can also be used for the actual access to resources. Grid middleware technologies are discussed in Subsection 3.1.3.

Furthermore, as outlined in this subsection, in the literature there are several different environments associated to the 'virtual laboratory' paradigm, varying from a very simple environment used for teaching to a complex environment used for remote (virtual) experimentations that bring several organizations together to share their resources. In the remaining of this thesis, the term virtual laboratory (VL) is used to refer to its broad definition. The Collaborative Experimentation Environment (CEE) later addressed in Chapter 4 of this thesis, however, refers to the broadest definition of the VL with an emphasis on supporting joint multi-disciplinary projects and collaboration, specifically information sharing among organizations and scientists. As such,

Figure 1.1: Collaborative experimentation environment and its relation to other support environments
it shall provide an integrated environment aggregating the functionalities provided by other different types of virtual laboratories (see Figure 1.1).

1.3 Research Objectives

A proper handling of all previously mentioned difficulties for e-science experiments is important and necessary to help scientists cope with the emerging challenging characteristics of their experiments. Although virtual laboratory approach is a promising starting point, there are still many aspects that need further research. One of the challenging aspects is the management of scientific information in a VL environment.

Information management, which constitutes the main focus of this thesis, plays an essential role in supporting scientific experiments, in the sense that it is involved in all types of e-science experiments and at any and all stages of the procedures followed during the experiments. In order to address the information management aspects of a virtual laboratory environment, the key research question is formulated as follows:

\textit{Is it possible to provide a flexible and extensible, yet robust information management platform that offers generic and reusable solutions to fulfill the information management requirements of emerging experiments and applications in different e-science domains?}

Consequently, the general objective of this thesis work is formulated as follows:

\textit{Analysis of the information management requirements of scientific experimentations in different e-science domains, and propose generic and reusable solutions for the management of scientific information in a collaborative virtual laboratory environment.}

The proposed information management platform will have to provide necessary mechanisms for proper handling of the information generated by e-science experiments. In specific, the information management platform will be responsible for:

- managing all information handled in the VL and stored in internal data resources, where the information can be scientific information generated by VL users, experiments and applications (e.g. experiment designs), or it can be administration or management-related information generated by the VL itself for its proper operation (e.g. user profiles);

- providing mechanisms for accessing both internal VL information resources (e.g. databases storing application-specific information) and external information resources (e.g. public repositories for scientific information), and managing accesses to these resources;

- handling information management requests of other (architectural) VL components as well as requests of VL users and applications.
In order to offer complete solutions to scientists, the information management platform must be an *integral* constituent of the virtual laboratory environment (see Figure 1.2). As shown in this figure, VL users and applications access and use data/information resources through the VL information management platform, and hardware/software resources through other VL components or through a resource management middleware. That means, VL users and applications access and use all resources they need through a single VL environment. Such an integrated approach will eventually pave the way to achieving *one-stop-experimentation*, where scientists will perform all their scientific activities in this single VL environment.

Among the various types of data/information generated by e-science experiments, some data sets are very large in size and unstructured, and their analysis do not require any specific Database Management System (DBMS) functionality (e.g. querying functionality). These data sets can be analyzed by high-throughput software, hence, for performance reasons mainly, they may not be stored in databases that are managed by DBMSs. In such cases, high-performance storage facilities are used instead of a DBMS for the storage of large data sets as files (please refer to ‘data characterization’ in Section 2.2 for further details). Note in Figure 1.2 that applications and other VL components can access these high-performance storage resources via a resource management middleware (dashed arrows). Several ongoing research efforts are focused on the storage and manipulation of large data sets, especially using the

![Figure 1.2: Accessing data/information resources in a VL through its information management platform](image-url)
Grid middleware technology (see Subsection 3.1.3 for an overview of these efforts). Hence, file manipulation facilities provided by the Grid are considered in this thesis for the management of very large data sets that are stored in files.

In this direction, the specific objectives of the work described in this thesis, thus generating the thesis contribution, are as follows:

- Designing and developing the information management platform as an integral part of a virtual laboratory environment.
- Supporting the transfer of knowledge and expertise among scientists by enabling the sharing and reuse of experimental designs and information.
- Modelling different aspects of the information management platform, including the modelling of experiments, functionality, and (experiment and information) representation in a generic form to be used by different scientific domains and applications; as such supporting multi-disciplinary scientific research.
- Providing generic and reusable mechanisms for storage and manipulation of heterogeneous scientific information and complex data and information types required, generated, and manipulated during an experiment.
- Providing mechanisms for secure/authorized information sharing and exchange among users and application domains.
- Enabling access to multiple data/information resources to support collaboration.
- Designing and developing a scalable environment that is open for adding new users and resources.
- Providing a flexible and extensible environment that is open for adding new types of information.
- Designing and developing a highly modular platform that is easy to maintain and upgrade.
- Using existing and emerging standards, as applicable.
- Defining a methodology for integrating new scientific applications into the developed information management platform.

1.4 Research Approach

This section provides a summary of the approach applied to the research presented in this dissertation. Considering the multi-disciplinary nature of the field, the research has been performed through a number of phases. Figure 1.3 shows different phases followed in this approach, and the main results of each phase that are used as input for the next phase. Each phase is described below:
1. **Field study.** The first phase of the research approach was to make a comprehensive study of two representative e-science application cases: The *DNA microarray application* from the biology domain and the *material analysis application* from the physics domain. During this study, a number of experimental procedures that are used for (different parts of) experiments have been identified and studied. Step-wise descriptions of these procedures have been generated (i.e. description of each step in the procedures). Furthermore, for the steps that correspond to activities, protocols used for the accomplishment of the activity, hardware and software tools and their required parameters were all identified and defined. This was a very detailed study of the experiment procedures, that resulted in a fine-grained characterization of different types of experiments performed in these two application cases. Detail granularity for these two case studies were down to the level of the properties of a single step in each procedure. Besides these two main application cases, a few other scientific/engineering applications were also studied, including the confocal microscopy experiments and traffic simulation applications. All these studies were done by visiting laboratories, attending demonstrations, and many face-to-face discussions with the scientists and technicians from these different domains.

Following this work, a number of similar applications were studied from the literature, and a long list of difficulties that scientists face have been identified. The refinement and analysis of all these study results were focused on identification of information management requirements.
2. **Related research.** Considering the main goal of this dissertation on provision of a supporting environment for scientific experimentations, the second phase consisted of an extensive research of the related work. This related work study encompassed both research on technologies, paradigms and standards that enable the development of a supporting information management platform, and identification of the other innovative proposed solutions addressing the management of scientific information. The focus here was to see how other solutions do/do not address the long list of requirements that were identified in the previous study phase.

3. **Designing a solution.** This phase of research involved the modelling, design, and system implementation activities. Since the VL shall provide a uniform experimentation environment for multi-disciplinary research, main focus during this phase was on the **generalization** and **reusability**.

   First, a data model for information about scientific experimentations has been developed for supporting the two main application cases. Then this model has been generalized, also by using the results from the study of other applications as input. The generalized model supports the modelling of information about scientific experimentations from different e-science domains. At this phase of research, the complexity and diverse characteristics of the two initial application cases played an important role in the achievement of such generalization. At this stage, in parallel to the information modelling, concepts and (usage and presentation) models were also generalized. A generic experiment concept has been defined, and a generic (usage and presentation) model for user environment has been developed.

   The generalization of the models and concepts allowed the design of an information management architecture that is capable of uniformly managing information about scientific experiments. Furthermore, the generic model for the user environment allowed the design of the user interfaces that uniformly present the experiments and information about them to scientists. On the other hand, special attention was paid on the modularity of the architecture during the design phase, since modularity is among the key aspects of reusable software components.

   Finally this architecture was implemented, by applying as much as possible standards and open source tools and technologies as applicable.

4. **Validation of the proposed solution.** When the development of this information management platform was completed, the (generalized) concepts and models have been re-applied to the two initial application cases. Necessary changes and improvements have been identified; and the models, concepts, architecture and implementation have been updated accordingly for validation and verification of these modifications. In addition, during this validation phase further generalization took place, which allowed replacing some of the specific implementation details with more general (and reusable) counterparts, mainly in further generalization of the service interfaces.
1.5 Base Projects

This section presents an overview of the VLAM-G project that formed the base for this thesis work, and the other projects that contributed.

1.5.1 The VLAM-G Project

The Dutch ICES/KIS-II project VLAM-G – Grid-based Virtual Laboratory Amsterdam [60, 68, 69] provides the main context for the research work presented in this thesis.

The VLAM-G is a multi-disciplinary virtual laboratory environment for remote experiment control and Grid-based distributed analysis in experimental sciences, using cross-institutional integration of various resources. By providing the necessary Grid-based distributed high-performance hardware and software infrastructure, it enables scientists and engineers in different areas of research to work on their problems via experimentation, while making optimum use of the modern information technology. Although the initial focus was on the application cases from the domains of biology, physics and systems engineering, the VLAM-G provides the required generic environment for multi-disciplinary research in all e-science domains.

The VLAM-G virtual laboratory environment allows its users to:

- perform multi-disciplinary, collaborative experiments in a uniform, integrated environment,
- complement their in-vitro experiments with in-silico experiments,
- combine data integration with software integration,
- define customized experimental procedures and analysis flows,
- reuse generic software components, and
- share hardware, software, storage, networking resources as well as knowledge and experience.

From a higher-level point of view, the VLAM-G acts as a bridge between the high-level user environments and the low-level Grid facilities (Figure 1.4). It hides the low-level details of the Grid layer by providing higher-level functionalities to its users to transparently access the Grid facilities.

The VLAM-G Architecture

The four-tier architecture of the VLAM-G is shown in Figure 1.5. Application presentation tier includes the user environments which allow users to interact with the VLAM-G and use its facilities. Components in the application toolkit tier provides the generic VL functionality, on top which application-specific functionality can be developed. Grid services tier provides the distributed resource management functionality using the Grid middleware technology. VLAM-G resources tier consists of
the computing, networking, and storage resources that are made available to users and applications through the VLAM-G. Figure 1.5 also shows the components of the VLAM-G architecture distributed over these four tiers, while interaction among the various components of the architecture is depicted in Figure 1.6. Seven main components of the architecture can be distinguished: Front-End, Session Manager, RTS, Collaboration System, Assistant, Module Repository, and Information Management (VIMCO). Below, each of these components is briefly described.

**Front-End.** *Front-End* is the user environment of the VLAM-G. Users interact with the VLAM-G only through the Front-End. The Front-End consists of a number of GUIs, which present the VLAM-G functionality to its users in a uniform way.

**Session Manager.** *Session Manager* [70] acts as a gateway to the VLAM-G server side. It manages the active user sessions, and is responsible for coordinating the interactions among the VLAM-G components. All user requests originating from the Front-End first arrives to the Session Manager, which then dispatches/forwards the requests to the corresponding VLAM-G component.

**Run Time System (RTS).** The distributed computing and networking resources on the Grid are made available to VLAM-G users through the *Run Time System* [71]. It makes use of the Globus [72] services, while hiding the details of this low-level toolkit from users. Generic VLAM-G processing functionality and application-specific problem solvers are presented to its users as self-contained modules. RTS provides a module API which encapsulates the Grid computing

---

**Figure 1.4:** Overview of the VLAM-G as a bridge between high-level user environments and low-level Grid facilities
code within a simple module interface. Users compose their experiments by attaching a number of RTS modules to each other, which are then executed by the RTS in the Grid environment.

Collaboration System & Assistant. Collaboration component of the VLAM-G will enable simultaneous collaborative design and execution of experiments. The collaboration system will offer audio and video communication between participating scientists by supporting among others a whiteboard for scientists. The Assistant, on the other hand, is a subsystem that will assist users during the design of an experiment, for instance, by suggesting a user the most-efficient module to perform a specific task. Implementation of these two components is planned for future work.

Module Repository. Module Repository is a persistent storage for the executable binaries of VLAM-G software modules.

VIMCO. The VIMCO – Virtual Laboratory Information Management for Cooperation is the information management component of the VLAM-G architecture.

The DNA microarray and the MACS applications were the two initial application cases of VLAM-G. The work in this thesis has been initiated and motivated by the need to support the information management requirements of both these initial application cases and any future applications within the VLAM-G.

Figure 1.5: Four-tier architecture of the VLAM-G and its components
VLAM-G Development

VLAM-G is a joint effort involving several partners from different disciplines. The work presented in this thesis addresses the information management aspects of the VLAM-G environment. However, since the information management forms the base for the interactions among the VLAM-G components as well as the interactions between the VLAM-G and its users, the design of the information management platform had also consequences on the VLAM-G components other than the information management. As a result, the author was involved in the design activities of the other VLAM-G components, especially in the modelling of the VLAM-G Front-End, as presented in Chapters 4 and 5. However, the architectural design and the actual implementation of the Front-End were realized by other VLAM-G members.

1.5.2 Other Related Projects

Besides the VLAM-G – VIMCO project, the author has been involved in many other projects as a member of the Cooperative Information Management (CO-IM) Group [73] of the University of Amsterdam (UvA), which directly or indirectly contributed to the work described in this thesis. These projects include: Dutch ICES/KIS VLAM-G – MACS Project (1999-2002), Dutch NWO-BMI Flexwork Project (2001-2004),

![Diagram of VLAM-G architecture](image)

Figure 1.6: Components of the VLAM-G architecture and their interaction

1.6 Structure of the Thesis

This dissertation describes the results of the research performed towards the design and development of a scientific information management platform for collaborative experimentation environments.

The structure of the thesis conforms to the phases of the research approach that was presented in Section 1.4. Results of each phase are presented in one or more chapters of the thesis, as follows:

Field study: Chapter 1 – Characterization of e-science domains and experiments/applications  
Chapter 2 – Extensive characterization of VL, use case analysis, requirements analysis

Related research: Chapter 3 – Study, evaluation, and comparison of related systems

Design: Chapter 4 – Design of the solution data models and functionality, definition of a methodology for extension into other scientific domains and applications  
Chapter 5 – Design and implementation of the software platform, based on the data models and functionality presented in Chapter 4

Validation: Chapter 6 – Validation of the results of Chapters 4 and 5 through extensive description of their application to real world case studies

The main topics elaborated in each chapter are summarized below.

Chapter 1 described the emerging e-science paradigm, defined and characterized e-science domains and applications, and provided as motivation examples of emerging scientific experiments to illustrate the difficulties that scientists face during their experimentations. Next, the virtual laboratory solution was introduced. This chapter then presented the objectives of this thesis work and the research approach applied to achieve these objectives. The projects that provided the context for the research work and contributed to the results obtained were also described in this chapter.

Chapter 2 first provides a characterization of the proposed virtual laboratory solution, then focuses on the analysis of the requirements from collaborative experimentation environments to properly address the identified problems in scientific information management. In this direction, this chapter presents the results of the use-case analysis performed to identify the different users of a VL environment, and outlines the needs and expectations of each user type as well as the requirements from the base ICT infrastructure to properly address the user needs.
Chapter 3 presents a framework for the study and evaluation of related work in support environments for scientific experimentations. This chapter first introduces the technologies, paradigms and standards that constitute the enablers for developing support environments. Remaining of this chapter describes the evaluation criteria for related work, provides summaries of the studied systems, and finally provides an evaluation of the studied systems with respect to the defined criteria.

Chapter 4 defines a framework for information management in CEEs. The chapter first addresses the modelling of scientific experiments which provides the base for the proper description of the other aspects of the framework, and describes a framework for CEE user environments. This chapter finally focuses both on data modelling for different types of experiment-related information handled in the CEE and on functionality modelling for the required functionality for proper management of this data.

Chapter 5 describes the implementation of different VLAM-G components, with the main focus on its information management platform (VIMCO). In this direction, first the experiment model adopted by the VLAM-G is described, followed by a brief overview of the VLAM-G Front-End. This chapter then focuses on VIMCO and describes its architecture as well as the databases it maintains and the services it provides.

The DNA microarray application case is presented in Chapter 6. Besides providing a brief overview of microarray experiments, this chapter further presents the Expressive System that is designed and developed to support the information management requirements of microarray experiments within the VLAM-G environment. The DNA microarray application case provides an example of how the developed information management platform can be applied to supporting specific application domains.

Finally, Chapter 7 summarizes the proposed approach, outlines the main achievements, evaluates the proposed VLAM-G/VIMCO approach and compares it with the related systems. This chapter then presents the ideas on future work and possible advance extensions.