Project no. 004260

DILIGENT

"A Testbed DIgital Library Infrastructure on Grid ENabled Technology"

Integrated Project

FP6-2003-IST-2

Publishable Final Activity Report

Period covered: from 1 September 2006 to 30 November 2007

Date of preparation: 28/01/08

Start date of project: 1 September 2004

Duration: 39 months

Coordinator:
Jessica Michel
ERCIM

Version 0

Scientific coordination:
Donatella Castelli
CNR-ISTI

1 Project execution

The main objective of DILIGENT has been **to create an advanced testbed knowledge e-Infrastructure enabling members of dynamic virtual e-Science organisations to access shared knowledge and to collaborate in a secure, coordinated, dynamic and cost-effective way**. This e-Infrastructure has been built by integrating Grid and Digital Library technologies. The merging of these two different technologies has opened the way to a next generation of e-Science knowledge e-Infrastructures able to provide powerful environments for research and industrial applications.

The DILIGENT e-Infrastructure has been experimented and validated by building two virtual research environments (VREs) serving complementary real-life application scenarios: one from the cultural heritage domain and one from the environmental e-Science domain. The first VRE has demonstrated how the DILIGENT infrastructure, by simplifying the processing of multimedia artefacts and offering support for collaborative multidisciplinary studies, can favour faster progress in research and better quality of education, also in research areas, like the humanities, traditionally far from the most advanced computing technologies. The second VRE has shown how a knowledge e-Infrastructure, by integrating distributed multi-type data sources with specialised data handling services, can improve accessibility, interoperability and usability of environmental data, models, tools and instruments.

Additional important objectives of the DILIGENT project have been:

- to open up grid technology to the broader range of research and industrial communities which produce, process and consume knowledge;
- to broaden the diffusion of digital libraries by providing a more cost-effective digital library creation and operational model, and by supporting new types of documents and tools able to satisfy the communication needs of a larger number of application areas;
- to promote cross-fertilisation between the digital libraries and grid technology domains, which can foster synergies and advances in both of these areas.

Fifteen contractors participated in the project activities:

1. European Research Consortium for Informatics and Mathematics (ERCIM, France)
2. Consiglio Nazionale delle Ricerche (CNR-ISTI, Italy)
3. National and Kapodestrian University of Athens (UoA, Greece)
4. Swiss Federal Institute of Technology Zurich (ETH, Switzerland; participation terminated 31 March 2005)
5. Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V. (FhG, Germany)
6. University for Health Informatics and Technology Tyrol (UMIT, Austria; participation terminated 31 March 2006)
7. European Organization for Nuclear Research (CERN, Switzerland)
8. Engineering Ingegneria Informatica SpA (ENG, Italy)
9. University of Strathclyde (USG, United Kingdom)
10. Fast Search & Transfer ASA (FAST, Norway)
11. European Space Agency (ESA-ESRIN, Italy)
12. Scuola Normale Superiore (SNS, Italy)
13. 4D SOFT Software Development Ltd. (4D-Soft, Hungary)
14. RAI Radio Televisione Italiana (RAI, Italy)
15. Universität Basel (UNIBAS, Switzerland)

The project management was jointly carried out by two individuals:

- Jessica Michel - Administrative and Financial Coordinator, and contact point to the European Community
  ERCIM EEIG (European Research Consortium for Informatics and Mathematics, a European Economic Interest Grouping)
  2004 route des Lucioles - BP93
  F-06902 Sophia Antipolis Cedex, France
  Tel: +33 4 9238 5089 e-Mail: Jessica.Michel@ercim.org
- Donatella Castelli – Technical and Scientific Co-ordinator
  CNR-ISTI (Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione)
  Area di Ricerca di Pisa
  Via Moruzzi, 1
  56124 Pisa, Italy
  Tel: +39 050 315 2902 e-Mail: Donatella.Castelli@isti.cnr.it

The work performed in the three years of the project lifetime was articulated in four areas that correspond to the main project objectives (see Figure 1):

1. **Technology integration**
2. **Experimentation with virtual user communities**
3. **Feedback to the constituent technologies**
4. **Exploitation and Sustainability**

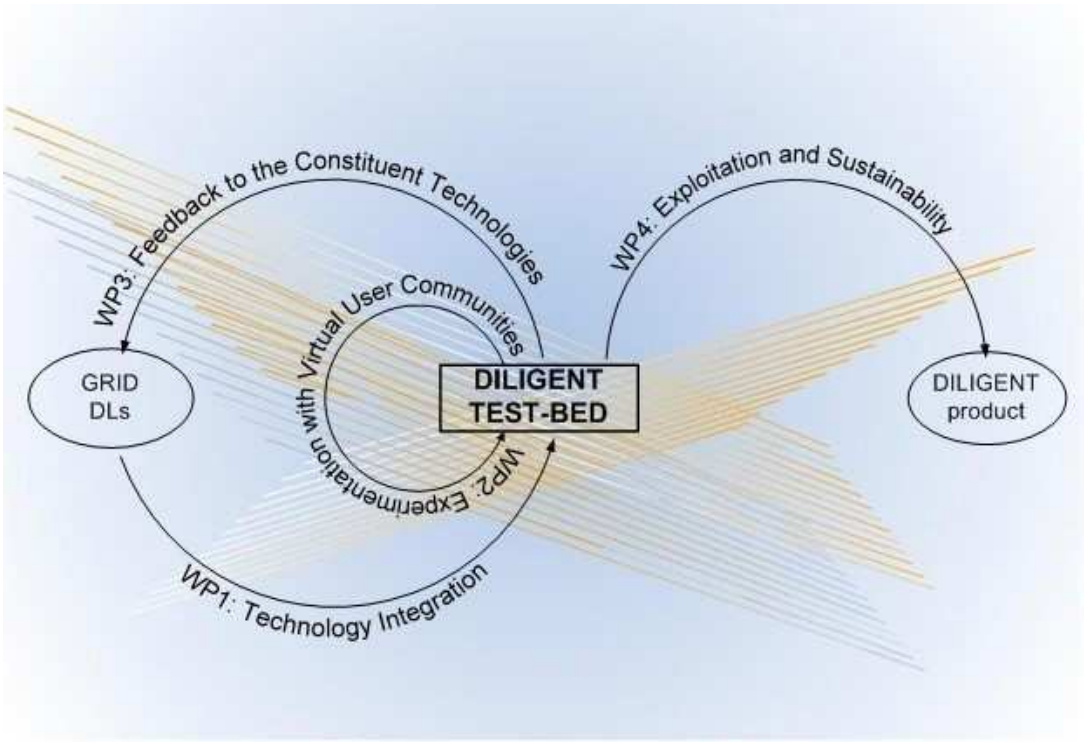

Figure 1: The DILIGENT project activity areas

A considerable part of the project activity was dedicated to design and develop a system named **gCube** (http://www.gcube-system.org). This system is a novel three-dimensional implementation of the resource sharing vision of the grid, entirely built on Java, which adopts gLite as the

underlying grid middleware and exploits several Globus Toolkit 4 elements. In particular, it provides:

- a feature-full platform for distributed hosting, management, processing, and retrieval of content;
- a framework, named gCube Core, for building state-of-the-art domain specific applications out of extendible building blocks, which opens unique opportunities for composition and outsourcing of implementations;
- a service-based grid middleware for cost-effective utilisation of computational and storage resources of grid infrastructures, that eliminates manual deployment overheads.

gCube is a Service Oriented System, logically partitioned in three layers, in turn organised in five functionally focused service groups (see Figure 2) that enable access to the infrastructure resources and offer the tools for building feature-rich end-user applications.

- *Collective Layer*

  Lays the foundations for the system by enhancing existing Grid collective services in order to support the complex service interactions required by the Information Management Layer;

- *Information Management Layer*

  Supports the storage, handling, and retrieval of multi-type and mixed-media content;

- *Application-Specific Layer*

  Gathers general-purpose application tools as well as APIs and SDKs for third parties to migrate their data or functional components to a gCube-powered infrastructure. Furthermore it builds a number of tools of common interest to domain-specific application developers.

The final gCube version released by the DILIGENT project consists of 137 components among services, libraries and portlets. All gCube services are WSRF and Web Service Notification compliant.

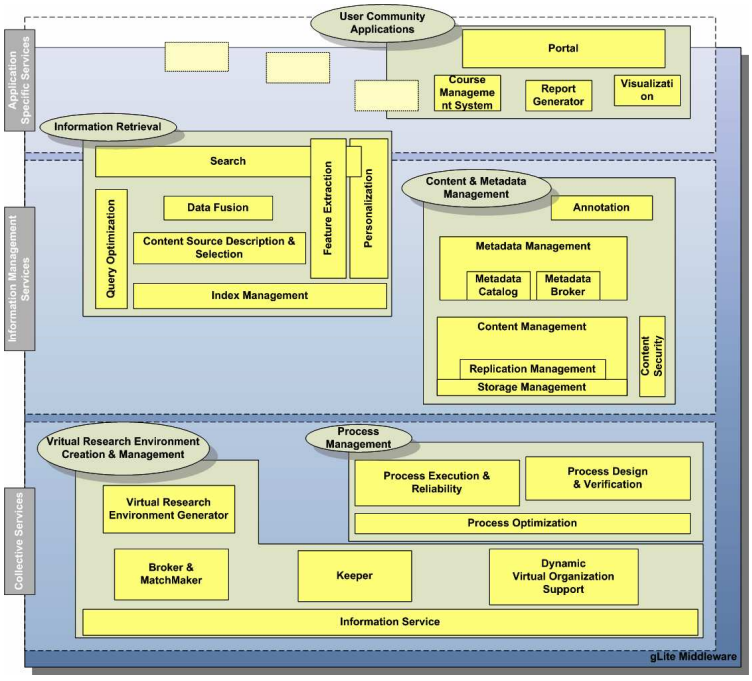

Figure 2 - The gCube logical architecture

Detailed manuals have been produced for developers, system administrators and end-users that illustrate how to use, extend and deploy the system and the created VREs.

A development infrastructure and a testing infrastructure were set up during the first period of the project. These infrastructures were exclusively based on DILIGENT resources. In the last year of the project a real pre-production infrastructure providing reliable gLite and gCube services was released. It consists of six DILIGENT gLite sites, five of them certified as EGEE PPS sites.

The availability of a pre-production infrastructure made it possible not only to support the planned experimentation scenarios but also to serve the needs of the EU co-funded SAPIR project (http://www.sapir.eu). SAPIR's goal is to provide searches over huge quantities of multimedia objects. Such searches are based on both text and multimedia features and exploit similarity and query-by-examples searches. SAPIR found in the DILIGENT infrastructure an extremely useful means to create the large collection of multimedia features it needed. A data challenge was then designed which aimed at executing feature extraction on images downloaded from the Flickr database (http://www.flickr.com). As part of this data challenge 1000 jobs were submitted per day (through 2 WMSs). Each job processed between 500 and 1000 images and required approximately 50 Mb of disk space and at least 512 of RAM. Jobs consumed between 20 minutes and 1 hour of CPU time (depending on CPU). As a result of the first phase of the experimentation 112 million products were generated that correspond to 4.55 TB of data and contain approximately 150 million features. Given the large interest raised by this experimentation and the interesting results obtained, a plan has been made to extend it also beyond the end of the DILIGENT project.

The two user communities represented in the consortium strictly collaborated with the DILIGENT technicians in order to **specify requirements for virtual research environments serving their own needs, provide content and applications to be fed in the DILIGENT infrastructure for experimentation and give feedback to the technicians about the products developed.**

In particular, the culture-heritage community, named **ARTE** (Applicazione di Ricerche e Tecnologie di Editoria Digitale) provided requirements for a VRE dedicated to support the collaborative creation of courses on specific topic of interest for a community composed by scholars of different Institutions. These scholars work together to set up the basis for a new research discipline that merges experiences from the humanities, social science, and communication research areas. This VRE, accessible through the portal hosting all the VREs (http://diligent.isti.cnr.it) (see Figure 3 - The ARTE VRE portal) provides access to a number of thematic authoritative and precious archives (literature, modern and ancient historical events, human heritage) and to state-of-the-art Information Retrieval techniques for the creation of dynamic learning material. The formats of the objects maintained in these archives are very heterogeneous, as they range from metadata to content, spanning from standard TEI encoded files to proprietary database formats, JPEG files and AVI encapsulated movies. Some of these archives belong to the DILIGENT partners (i.e. Scuola Normale Superiore and RAI), others to collaborating Institutions (e.g. Boscan collection, Emblem Project collection, Medita Video collection) and, finally, some are open access resources (e.g. OAI-PMH compliant ones). Through the ARTE VRE scientists retrieve information pertinent to the subject of their research and re-organize it in a variety of new forms: through a Workspace Management tool, for personal re-use, or through a Course Management tool, for publication. Courses are delivered through the Moodle system. This well known and fully fledged course management system has been fully integrated with the DILIGENT specific digital library services. Users of the ARTE VRE can create courses by accessing it either through the DILIGENT integrated interface or through the well known Moodle one. This secure and quite comfortable VRE further promotes the objectives

of the DILIGENT project by attracting communities and resources to the grid and achieves additional means for its long term sustainability. Until now ARTE users have created five courses through this VRE: “Con parola brieve e con figura”, “Emblems and devices from 16th and 17th century books”, “Astrology: Art and Culture in the Renaissance”, “The Art of Memory” and “Cupid Representation”. The creation and usage of these courses have provided very useful feedback that has enabled further tuning of the gCube functionality.

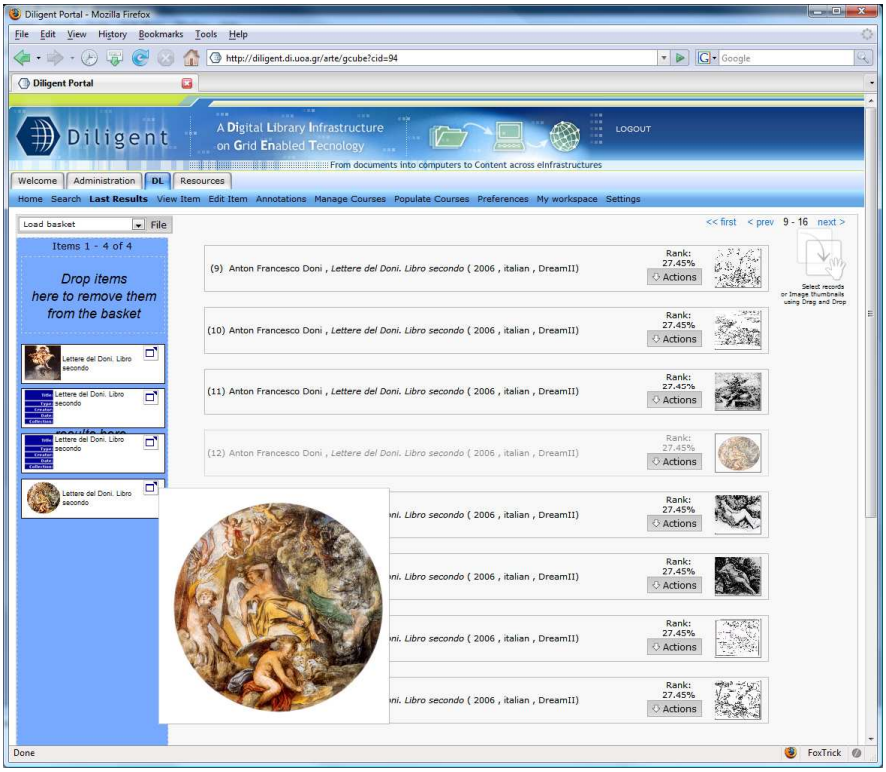

Figure 3 - The ARTE VRE portal

The environmental science community, named **ImpECt** (Implementation of Environmental Conventions), which includes leading actors in the environmental sector like the European Space Agency (ESA), proposed the creation of a VRE dedicated to support the generation of environmental reports. International and regional, continuously evolving and thematically specialized conventions related to marine pollution and also the UNESCO World Heritage Programme, represent the framework for formulating international environmental agreements. In this complex scenario, the information sources and the tools for analysing and processing data extracted from such sources to produce new information are dispersed among environmental agencies.

This scenario has been supported through an Earth Observation VRE providing access to several heterogeneous information collections made available in the DILIGENT infrastructure (http://diligent.isti.cnr.it) (see Figure 4). This VRE includes collections of environmental reports from UNEP-MAP, EEA and FAO and satellite imagery and maps collections from ESA, NASA, NOAA, and FAO GeoNetwork. Tools for on-demand processing of such data have also made available through the infrastructure and consequently through the VRE. In particular, several services available from the ESA Grid Processing On-Demand (GPOD, http://eogrid.esrin.esa.int) have been integrated by developing specific Web interfaces on them.

Figure 4 - The ImpECt VRE portal

The creation of an appropriate environment for this application scenario highlighted important requirements. The need of some level of semantic awareness to better manage earth observation data and to improve search capabilities emerged as one of the most recurring ones. In order to satisfy these requirements, in the last part of the project an activity dedicated to add a first level of such semantic awareness was initiated which was aimed at integrating and exploiting in gCube knowledge representation techniques together with annotations and ontology management. An existing prototypical tool for enabling ontology based annotation was integrated in the final release of gCube. This tool was customised by importing a structured ontology from NASA. In an evolution perspective of the gCube middleware, the experience done, though still under investigation, has produced many useful insights on how to further extend the semantic support.

The DILIGENT project has also performed a number of activities addressed to **publicly disseminate the outcomes of the project**, and to **define the best strategies for the exploitation of these outcomes**. These activities comprise a large number of training sessions, a variety of dissemination actions, and the production of a market study and a business plan as preliminary steps towards sustainability.

Training activities played a key role in the project. Several training courses and documentation material were produced through which the DILIGENT partners could transfer the acquired knowledge to the other members of the consortium and to interested groups outside the project. At the same time, various user communities became acquainted with the new functionality offered by the DILIGENT infrastructure and were trained on how to exploit it in their application context. A **digital library** was created to support the training activity (http://diligent-training.isti.cnr.it). This digital library contains a rich variety of documentation, like reports, PowerPoint presentations, videos of lectures and tutorials on the DILIGENT project and on related technologies - e.g., Web Services, Service-Oriented Architectures, gLite and other Grid technologies that, being rated as relevant by some partner, were recommended to the others. This documentation, which has grown considerably in the last years, can hardly be retrieved and used

today since it is spread on many different sites and it is organised according to heterogeneous set of criteria that are meaningful only for the communities that make them available, e.g. project work packages. The DILIGENT digital library overcomes this problem by presenting a homogeneous interface through which all these documents can be retrieved. In November 2007 the DILIGENT training digital library contained more than 10,000 documents (organised into four publicly accessible collection classes: Projects; Scientific papers; Standardization bodies; and Technical documentation) occupying more than 6 GB.

The design and realisation of a number of on-line **demonstrators** have also been performed to further contribute to the transfer of DILIGENT outcomes. Developed demonstrators have addressed three target user categories: developers, decision-makers ad end-users. Taking into account interests and skills of these categories, six demonstrators have been implemented to show users advantages of using the grid in the framework of typical digital library applications. Each demonstrator has been developed as a separate application and made available to users in a dedicated section of the DILIGENT project website (http://www.diligentproject.org/demonstrators).

DILIGENT's **dissemination strategy** has been centered on two major classes of activities: pure dissemination activities and liaison activities. The former includes publicity activities and scientific announcements. The main objective of these activities has been to get the project and its results known to all potentially interested communities and to relevant task forces in the fields of Digital Libraries (DL), Information Retrieval (IR), Long Term Data Preservation and Grid Computing. Traditional instruments and modern e-Tools have been developed and used to support them. In particular, a project Website (http://www.diligentproject.org) has been set up at the start-up of the project to give access to information related to the project, and a gCube dedicated Web site (http://www.gcube-system.org) has been publicly open when the Beta release of this system was delivered. More than forty dissemination events have been carried out at major EU and International conferences, events and open meetings and many scientific papers have been published that describe the results of the project.
Liaison activities aimed at creating collaboration links with other projects, task forces or user groups were organized so that resources could best be utilized under a global perspective. Liaison activities included commonly organized activities as well as exchange of technology and knowledge among collaborating parties.

In tight collaboration with the exploitation activities of the project, dissemination ones have also supported relations with the users of the DILIGENT infrastructure in order to promote sustainability.
We expect that these user communities can largely profit from the functionality implemented by this e-Infrastructure. The experience done during the DILIGENT project, that will be continued by the follow-up D4Science project (http://www.d4science.eu), will certainly result in the development of new technologies that better respond to the needs of these and many different research communities. The final expectation is that through the experiences done, it will be possible in a near future to build a new e-Infrastructure layer that by exploiting existing Géant and Grid infrastructures, will enable the sharing and creation of new knowledge resources at the European level and beyond, thus significantly transforming the way of conducting e-Science.

More information about the project can be found on the project Website: http://www.diligentproject.org

2 Dissemination and use

[Publishable results]

- Design patterns and abstract solutions for DL services designed to operate in a Service Oriented Architecture and capable to exploit Grid technologies;
- Experiences in developing an e-Infrastructure supporting a three dimensional notion of resource sharing (content, services and computing&storage capabilities) by exploiting state-of-the-art standards and technologies;
- Operational framework for supporting the implementation of dynamic VREs and aggregated applications on a Grid-based infrastructure; and
- VREs as operational paradigms for serving collaborative research activities.

[Descriptions]

1. **Design patterns and abstract solutions for DL services designed to operate in a Service Oriented Architecture and capable to exploit Grid technologies.** DILIGENT has been designed to operate in a novel scenario, where Digital Library and Grid technologies are combined to produce an integrated framework providing its users with the best of both worlds. In this scenario DILIGENT researchers had to face new issues as well as find new solutions for addressing them. The framework which resulted at the end of the project includes innovative solutions for, among other, VREs definition and dynamic deployment, workflows definition and execution, distributed information organisation and management, and distributed information retrieval. All these solutions have been designed with the goal to maximise impact and guarantee an optimal consumption of the available resources, ranging from computing and storage facilities to collections and services. A key aspect largely investigated in the project is the movement of large semi-structured (e.g., XML) data in a distributed system, which is especially difficult when partial access to the data is required.
2. **Experiences in developing an e-Infrastructure supporting a three dimensional notion of resource sharing (content, services and computing&storage capabilities) by exploiting state-of-the-art standards and technologies.**

   The initial goal of the project was to deliver a test-bed e-Infrastructure to prove the feasibility of the proposed solution. The DILIGENT project invested effort in realising a framework for solutions that could easily evolve in a fully-fledged foundation technology by exploiting a large amount of existing standards in order to deliver operational e-Infrastructure. In developing such a framework, a great deal of experience in combining existing standards was accumulated. The strengths and weaknesses of many existing standards and technologies were identified by concretely exploiting them in very challenging scenarios.

3. **Operational framework for supporting the implementation of dynamic VREs and aggregated applications on a Grid-based infrastructure**

   The controlled sharing mechanism underlying the notion of e-Infrastructure is a fundamental mechanism for implementing VREs and, in general, applications built by re-using existing assets. In particular, by promoting the three dimensional notion of sharing supported by DILIGENT and its enabling framework, i.e. gCube, the provision of shared resources and the construction of VREs and, more generally, of dynamic applications is largely simplified.

4. **VREs as operational paradigms for serving collaborative research activities.** Nowadays research activities are becoming very multidisciplinary and require virtual research organisations, i.e. dynamic organisations of researchers crossing the boundaries of single institutions. In addition, the members of such organisations have the need to establish strong collaboration channels and environments in a relatively short time interval and with a limited budget, i.e. many cannot afford the development of a system that will serve their needs. The VRE paradigm and support implemented by DILIGENT is a solution for such an application scenario, fitting the needs of several research disciplines.