Information Society Technologies

# Diligent

A DIgital LIbrary Infrastructure on Grid ENabled Technology

## Market and Technology Trends Analysis

**VERSION 1.2**

May 2, 2005

# Document Information

| | |
|---|---|
| Project | |
| Project Title: | DILIGENT, A **DI**gital **L**ibrary **I**nfrastructure on **Grid EN**abled **T**echnology |
| Project Start: | 1 Sep 2004 |
| Call/Instrument: | FP6-2003-IST-2/IP |
| Contract Number: | 004260 |
| Document | |
| Deliverable number: | 4.3.1 |
| Deliverable title: | Market and Technology Trends Analysis |
| Contractual Date of Delivery: | Month 6 |
| Actual Date of Delivery: | 30th April 2005 |
| Editor(s): | Ingvar A. Aaberg, Andrea Manieri |
| Author(s): | Ingvar A. Aaberg, Pedro Andrade, Maria Bruna Baldacci, Donatella Castelli, Peter Fankhauser, Andrea Manieri, Roberta Faggian Marque, Claudia Niederee, Francesco Saverio Nucci |
| Participant(s): | CERN, CNR-ISTI, ENG, FAST, FhG/IPSI |
| Workpackage: | WP4.3 |
| Workpackage title: | Exploitation |
| Workpackage leader: | ENG |
| Workpackage participants: | CERN, CNR-ISTI, ENG, FAST, FhG/IPSI |
| Est. Person-months: | 5 |
| Distribution: | Internal |
| Nature: | Report |
| Version/Revision: | 1.2 |
| Draft/Final | Final |
| Total number of pages: (including cover) | 84 |
| File name: | D4 3 1 Market and Technology Trends Analysis - V1.2.doc |
| Key words: | *Digital libraries, Grid, utility computing, e-Science, eu-ist* |

| **Document log** | | | |
|---|---|---|---|
| Version | Date | Comment | Author |
| 0.1 | 2005-02-02 | First TOC | Ingvar Aaberg |
| 0.2 | 2005-02-09 | Updated TOC | Ingvar Aaberg |
| 0.3 | 2005-02-25 | First draft with input from all | FAST, CERN, CNR, ENG |
| 0.4 | 2005-03-04 | Incorporated feedback on first draft | FAST, CERN, CNR, ENG |
| 0.5 | 2005-03-18 | Incorporated feedback on version 0.4 | FAST, CERN, CNR, ENG |
| 0.5.1 | 2005-03-22 | Early revision | Andrea Manieri |
| 0.6 | 2005-03-29 | Formal revision | Jessica Michel |
| 0.7 | 2005-03-31 | Reorganisation of TOC with the insertion of Executive Summary and conclusion chapters; inclusion of a description of the methodology used; improvement of the overall content of the doc. | Francesco Nucci, Andrea Manieri, Ingvar Aaberg |
| 0.8 | 2005-03-01 | Inclusion of CERN first contributions and CNR comments on the quality of some comments; final rationalisation of the whole doc; addition of a "disclaimer" section | Andrea Manieri, CERN, CNR |
| 0.9 | 2005-04-07 | Inclusion of FhG/IPSI, CERN contributions and overall revision of the doc | Andrea Manieri, FhG/IPSI, CERN, Francesco S. Nucci |
| 0.10 | 2005-04-08 | Inclusion of CNR contributions | Andrea Manieri, CNR, Francesco S. Nucci |
| 0.11 | 2005-04-12 | General review | Ingvar A. Aaberg |
| 1.1 | 2005-04-27 | Improvements based on SC comments | Andrea Manieri, Ingvar A. Aaberg |
| 1.2 | 2005-05-02 | Revised executive summary | Ingvar A. Aaberg |

# DISCLAIMER

This document contains description of the DILIGENT project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The contents of this publication are the sole responsibility of DILIGENT consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 25 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

**DILIGENT is a project partially funded by the European Union**

# 1. EXECUTIVE SUMMARY

This report contains the technology and market analysis which will form the foundation for developing DILIGENT's sustainability model in D4.3.2.

The document contains an analysis of what technologies are available in the Digital Libraries market, compares them with Grid technologies, analyzes Digital Libraries user requirements, and looks at the joint capabilities of DILIGENT's consortium members.

We expect relevant technologies and market requirements to evolve significantly during DILIGENT's duration. Therefore, a yearly analysis will be performed to secure up-to-date information on relevant market needs, technology developments, competition, and new business opportunities.

The DILIGENT solution is a reusable framework for dynamic creation of Grid-based Digital Libraries solutions and services for e-Science communities and organization. The framework allows Virtual Research Organisations to integrate archive and other services required to meet their users' needs.

DILIGENT's value proposition is to be a key component in the value chain which brings together content providers, service providers, and storage and computation providers in the delivery of Digital Libraries services.

The report shows how the demand for making vast amounts of information available and retrievable increasingly is gaining the attention of corporations and governments. Recently, Google – the US web search company, announced it will fund PrintGoogle which is a 150 – 150 million US-dollar project to digitalise the books of 5 US and UK libraries. As an apparent response to Google's initiative, President Jacques Chirac of France wants to establish a similar and EU-coordinated program for European texts and libraries.

No vendors of commercial Digital Libraries exist today, and most efforts in the field have been based on tailoring solutions to meet project and collection specific needs. DILIGENT's platform and service approach seems to be unique. No comprehensive standards addressing the broad needs of Digital Libraries exist.

Our report describes a number of projects and initiatives within the Digital Libraries and Virtual Research Organization space, including DSpace, Greenstone, OpenDLib, BSCW Server, Sun Developer Network, Wikipedia, etc. The functional, operational and management requirements of Digital Libraries and Virtual Research Organizations are analyzed and documented.

Further, the report identifies and describes key enabling technologies for Digital Libraries, including content management systems, document management systems, media asset management systems, knowledge management systems, and workflow management systems.

Based on a technology and market study, our analysis shows how Grid technology is evolving – primarily through US-based research projects. One significant exception to the US-based initiatives is the European EGEE project, which is in the process of developing

Grid middleware and creating a production Grid. DILIGENT will use EGEE's middleware and are also in discussions on deploying Digital Libraries services on top of the EGEE production Grid. Some commercial Grid providers exist, and major commercial players, such as Sun, Oracle, and IBM - are starting to use the term Grid for marketing purposes.

Grid computing has evolved from distributed computing, and possible substitute technologies are analyzed in this report.

In parallel with the evolution of Digital Libraries systems we see an increase in the demand for Digital Libraries. Research work today, in any field, is often a collaborative effort carried out by groups belonging to different organizations worldwide. Motivated by a common goal and funding opportunities, these groups dynamically form virtual research organizations that share resources e.g., knowledge, experimentation results, instruments, for the duration of their collaboration, creating new and more powerful virtual research environments. These virtual research organizations, set-up by individuals that do not necessarily have a great economic power and technical expertise more and more frequently require Digital Libraries as tools for accelerating their research activities.

An analysis of library usage across Europe shows an increasing need for libraries offering digital services to their users.

Using a broad range of characteristics, our analysis indicates that libraries, e-Science and media companies are the most attractive market segments for DILIGENT to focus on. The opportunity for Digital Libraries has been analyzed for the following markets: Research collaboration, life sciences, medicine, financial markets, engineering and design, collaborative games, media, and government.

In addition to this summary (Chapter 1), the document contains seven chapters that prepare the ground for developing the DILIGENT sustainability model in the D4.3.2 delivery. Chapter 2 describes our methodology and provides a description of the DILIGENT project. Chapter 3 documents DILIGENT's context by describing initiatives related to the creation and management of digital contents. Chapter 4 describes enabling technologies, Chapter 5 documents technology and market trends relevant to DILIGENT, and Chapter 6 describes the market for Digital Libraries. The DILIGENT capabilities chapter (Chapter 7) contains a brief summary of the joint capabilities of DILIGENT's consortium partners. Finally, Chapter 8 concludes the report with critical success factors and a SWOT analysis.

# Table of Contents

# 2. Background

## 2.1 A METHODIC APPROACH TO MARKET ANALYSIS

This document contains an analysis of the market and technology environments DILIGENT exists in. The purpose of the analysis is to lay the foundation for D4.3.2 “Initial Exploitation and Sustainability plan”, which will identify, select and prepare for the execution of DILIGENT’s sustainability strategy.

### 2.1.1 A three step analysis

The analysis has been conducted by developing a methodology of multi-step or multi-level analysis which consists of three different levels. For each level, information has been collected to produce new valuable knowledge for a preliminary definition of the potential market for DILIGENT services.

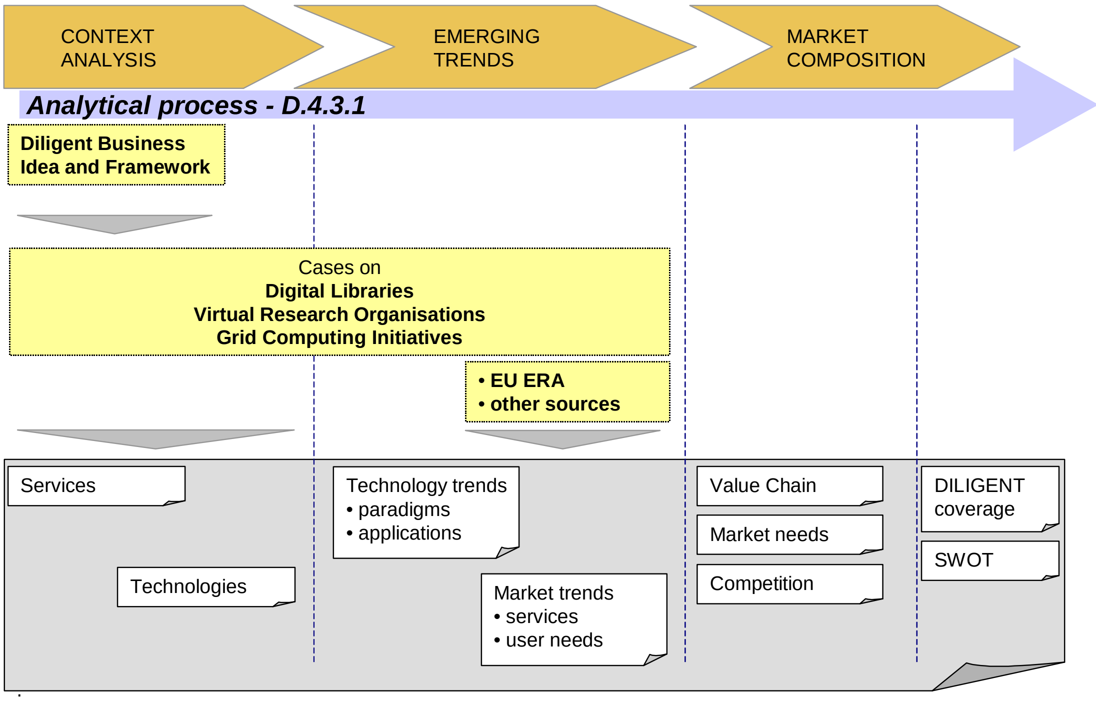
In particular, the context analysis aims to understand the international logical framework in which the project exists, focusing on the following aspects:

1. the **first aspects refers to the organisational perspectives** identifying analogous experience in Digital Libraries and initiatives supporting Virtual Communities (VCs), hence – in our context Virtual Research Organisations (VROs);

2. the **second issue refers to technologies enabling services** supplied by Digital Libraries. We have identified the use of CMS technologies as the common starting point for many DLMS (at least from a functional point of view). We have not found anyone attempting - like DILIGENT -to exploit the computing and storage power of Grid infrastructure to improve the functionality supplied by Digital Libraries;

3. Finally, the **market composition** applies all the considerations, technology and market insights, information and knowledge extracted, to identify the possible market addressable by DILIGENT, both in terms of services to offer and users or customers needs;

The following scheme illustrates the 3 different phases of the methodology, as well as the main section of the document.




### 2.1.2 An iterative process

DILIGENT is a three years project, over which we know that both relevant technologies and the market will evolve very quickly. Completion of generally agreed standards, new opportunities from related markets and new technologies can dramatically change the sustainability scenario for the project. Thus, a continuous and deep analysis of the markets and technology is part of the methodology applied to support the creation of a sustainability framework for DILIGENT.

The three-step process presented above will be re-applied every year in order to continuously have a clear vision of market needs, technology developments, competition, and new business opportunities. Such a deliverable belongs to WP4.3 Exploitation and Sustainability. Each partner has agreed to focus its attention on one specific issue and a partner will during its day-by-day activity collect and review relevant material (e.g., from conferences, Journals, books, projects, initiatives, etc.) in a structured form.

## 2.2 DILIGENT IN A NUTSHELL

### 2.2.1 Project objectives

The objective of the DILIGENT project is **to develop a digital library infrastructure testbed that facilitates knowledge sharing and remote co-operation in e-Science**. This infrastructure will support the on-demand creation of new generation Digital Libraries that will automatically be built by exploiting the set of shared resources maintained by the infrastructure itself. These resources are **content sources** (i.e., repositories of information searchable and accessible through a single “entrance”), **services** (i.e., software tools, that implement a specific functionality and whose descriptions, interfaces and bindings are defined and publicly available) and **hosting nodes** (i.e., networked entities that offer computing and storage capabilities and supply an environment for hosting content sources and services). By exploiting the functionality offered by this infrastructure multiple Virtual Research Organisations will be allowed to share their resources according to established policies and to build Digital Libraries that satisfy their specific co-operation needs.

### 2.2.2 Technical background

The DILIGENT infrastructure will be built by integrating two major and, thus far, independent technologies, Digital Libraries (DLs) and Grids. The idea to intertwine Grid and DLs will enhance not only these technologies per se, but also define systematic ways on how to develop next generation information networks on the basis of Grid technologies, which can be applied and adapted to other domains. The underlying idea is to promote the future generation of technologies in which computers and networks will be integrated into the everyday environment, rendering accessible a multitude of services and applications through easy-to-use human interface.

#### Digital libraries

Digital Libraries have been mainly intended as the digital counterpart to physical libraries. Since their introduction, they have been expected to provide mechanisms for maintaining distributed collections of documents, searching through these collections by exploiting the metadata records associated with the documents, and presenting the retrieved documents in a format suitable to the needs of the specific Digital Libraries audience. This understanding has been maintained for many years, even when Digital Libraries have been proved to be applicable not only to the library domain but to the entire Cultural field. Recently Digital Libraries have moved beyond the traditional connotation of the term “library”, and are rapidly shifting towards more general systems, now termed “dynamic universal knowledge environments”.

It is expected that these environments will correspond to the new generation of Digital Libraries. These Digital Libraries will be able to handle many multi-type object formats obtained through the combination of different multimedia components in an unlimited variety of ways. For example, they will be able to manage information objects that mix texts, scientific data and satellite images, or information objects that integrate images, annotations and videos. Also, the operations on these objects will be extensible in any direction without the limits imposed by the physical manifestation of the document. These operations, in turn, will be able to generate new information objects that may convey different semantic information. Furthermore, new Digital Libraries will be able to support the work of its users by providing functionalities that may range from general utilities, like annotation, summarization or co-operative work support, to user specific functions, like processing of maps, semantic analysis of images, simulation, etc. Through these new Digital Libraries, groups of individuals, which collaborate together to achieve a common goal, will be allowed to access, discuss and enhance the on-line shared information produced by them and by other groups or individuals. For example, in this environment a scientist will be enabled to annotate the article of a colleague with a programme that extracts useful information from a large amount of data collected by a specific observatory.

This annotation, executed on-demand when the annotation is accessed, will complement the content of the paper with continuously updated new information.

**Grid**

The increasing demand for intensive computation and processing of very large amounts of information (data, multimedia, documents, etc.) highlights the need to perform experiments at lower costs. At the same time the distributed and dynamic nature of scientific collaborations requires infrastructures and applications to be able to adapt to the needs of various user communities. In this context Grid computing has gained a lot of attention within the academic and scientific community and the IT industry.

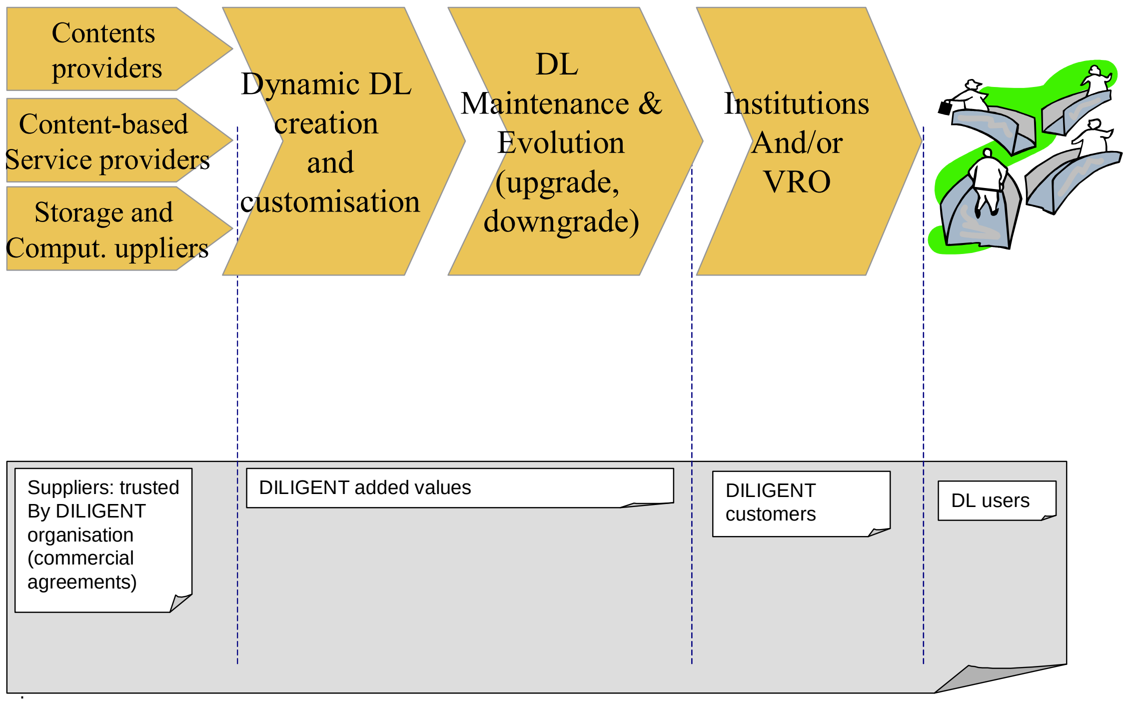
In the Grid vision, once a proper kind of infrastructure is in place, a user will have access to a “virtual computer” that is reliable and adaptable to the user’s needs. The resources - which may be processing power, storage, instruments, data or applications - will be managed dynamically and their allocation will be a transparent process.

Grid infrastructures have been implemented and experimented throughout the world in the last few years. Some have been demonstrated in operational environments, and on-going work is now aimed at creating a reliable and dependable European Grid Infrastructure that will integrate the results achieved in previous European and nationally funded projects. Grid technology has the potential to become a basic technology in e-Science.

### 2.2.3 The business idea

The DILIGENT solution is a framework for dynamic creation and maintenance of Digital Libraries for e-Science communities and organizations. Such a framework will improve the classic use of Digital Libraries by utilizing Grid technology. Using this framework any Virtual Research Organisation will be able to integrate any kind of archive and any third party services that could complement basic functions to accomplish with VRO needs.

DILIGENT organization will supply a service for the creation of **on-demand Dynamic Digital Libraries in a Pay-per-use fashion**, putting together resources eventually supplied by third parties. The DILIGENT service is related either with the creation of the Digital Libraries based on VRO needs and with the maintenance (and potentially the evolution) during the Digital Libraries lifetime.




In this vision DILIGENT creates a new value chain in the Digital Libraries business model, bringing together Content Providers, Service Providers and Storage and/or Computation Providers, to supply a new service to Institutions and Organisations (mainly virtual) that need to create Digital Libraries for their end-users. Examples of DILIGENT customers are, hence, public and private libraries, pools of federated libraries (maybe local libraries), temporary aggregation of institutions (e.g., crisis management teams, events organisation teams, etc), conference organisers, etc.

The DILIGENT business idea is based on the assumption that any content is available and accessible in some electronic form. At the moment such assumption is not completely true, since various initiatives on digitalization are still ongoing with the objective to digitalize European Heritage and to improve the techniques and technology for this task.

When this paper was edited, newspapers started writing about Google's PrintGoogle, where Google funds a 100-million-dollar project to digitalise the books of 5 US and English libraries. The press writes that this is the first step towards the so-called "Universal Library" where all the libraries content will be available by users for free.

As a response, President Jacques Chirac of France "… asked his culture minister, Renaud Donnedieu de Vabres, and Jean-Noël Jeanneney, head of France's Bibliothèque Nationale, to do the same for French texts—and create a home-grown search-engine to browse them." [Economist, April 2, 2005]

A deep analysis of this issue and it's implication on potential business and on sustainability models will be addressed in the deliverable D4.3.2 "initial business and sustainable plan" due at month 12 (due date August 2005).

# 3. The functional perspective

From a functional perspective, DILIGENT aims at addressing two set of requirements: one from the Digital Libraries environment and the other from the Virtual Research Organisations.

We have looked at systems and projects developing DLMSs (Digital Library Management Systems) in order to identify a common set of services for a DLMS. Since no commercial vendors of DLMS exist many projects deal with the creation of an ad-hoc DL. All these projects start from scratch creating systems handling already defined content sources. In our vision this duplication of effort and, sometimes, of content sources could be avoided: as with Data Base Management Systems (DBMSs) a set of basic common functions allows the easy creation of a Data Base (DB) and the reuse of archives, so with DLs there's the need to have a software that allow the easy creation (and maintenance) of a DL. Such software – as in the DILIGENT objectives – could be developed exploiting the Grids capabilities.

In addition, we have analysed projects supporting virtual collaboration and/or community. Even though DILIGENT deals with VROs a clear definition of a VRO doesn't exist. More in general people talk about Virtual Community as the *virtual aggregation of individuals with a common objective*. In the research environment this concept is misleading, since we recognise that some characteristics are less general. Our approach has been pragmatic: we studied projects where the two aspects of a VRO (i.e., the virtualisation of collaboration and the aspect to be a community) were considered to understand the minimum set of functionalities usually supplied to VCs, hence, possibly to VROs.

A selection of services to be supplied to final users by the DILIGENT solution is far beyond the scope of this deliverable. Such studies will be analysed in the light of user requirements collected in the project activities and compared with them in the design phase.

## 3.1 DIGITAL LIBRARIES SYSTEMS TODAY

Looking back at the history of the Digital Libraries field, we observe that almost all efforts to build such systems have been characterized by attempts to organize and provide access to particular collections of data and information, and to build systems tailored to the particular needs and characteristics of a specific target environment. As a result of this approach every development effort has been independent of any other and it has not exploited any of the already developed services. A lot of effort from experienced people has been required for preparing the digitized content and for developing the software that implements the Digital Libraries functionality. Dedicated computers, sometimes also quite powerful, as in the case of video Digital Libraries, had to be acquired in order to store and process the documents. These dedicated resources had to be dimensioned to support the highest peak of activities, even if these were executed only rarely, e.g., at start-up of the Digital Libraries or periodically for preservation purposes. As a consequence of this situation, Digital Libraries have been created only to serve large research communities or important institutions since these were the only ones that could afford the expenses of such products.

Currently, there are no established standards or even universally accepted architectural approaches for systems able to support both traditional and new Digital Libraries. So far the construction of Digital Libraries has been approached mostly in ad hoc ways. Moreover, there has been no effort to raise architectural issues to the forefront or compare alternatives to identify the areas of strengths and weaknesses for each one and arrive at a small number of dominant ones. This lack of standards implies high costs and long time of development and maintenance which restrict the exploitation of Digital Libraries only to restricted groups of potential user communities.

The understanding and expectation of Digital Libraries have evolved a lot since the nineties when the first digital library systems were built.

Many factors have contributed to this evolution, like the advent of new technological solutions in data communication, the introduction of new modalities in which information can be accessed and retrieved, the growth of the maturity of users that now demands richer and better quality services.

The initial Digital Libraries were mainly intended as the digital analogous of the physical libraries. They were providing mechanisms for maintaining distributed collections of documents, searching through these collections by exploiting the metadata records associated with the documents, and presenting the retrieved documents in a format suitable to the needs of the specific Digital Libraries audience. This understanding has been maintained for many years, even when Digital Libraries have been proved to be applicable not only to the library domain but to the entire Cultural field.

After approximately ten years of study and development it has now become clear that Digital Libraries systems can actually offer much richer functionalities than initially expected and that, if we will be able to enlarge their diffusion, they can transform the way in which joint research is conducted.

Recently, research projects have started to deliver practical basic systems for building Digital Libraries. Most of them are available as open source or as free services.

### 3.1.1 Cost of development

Setting up a Digital Library typically is a long term investment. It requires different types of complementing activities, which all contribute to the cost of entry for successfully exploiting digital library technology. This includes the set up and establishment of the respective IT system, the acquisition and integration of adequate content (also including metadata), and the effective implementation of organizational processes that ensure the targeted evolution and long-term operation of the digital library.

In the area of technology, the entry costs for setting up a digital library are similar to those of introducing other IT systems. Either a digital library system is implemented or existing system components that are selected, installed, integrated, and customized in order to implement the required digital library functionality. This implies costs for adequate hardware, licences and customization or complete development, training, etc. In addition, there are the typical operational costs like costs for maintenance and support.

In addition to the described technology-centred costs, there are the costs for "filling" the digital library with relevant content that makes it useful and attractive for the targeted community. This includes relevant content into the digital library and the import or generation of the required metadata to make the content accessible. The implied costs do not only depend on the availability of content (in digital form) and on the effort required for creating the metadata (import, semi-automatic generation or manual entry), but also on the openness and flexibility of the selected digital library systems.

Finally, organizational processes like collection strategies and methods for ensuring long-term accessibility and preservation have to be defined and implemented. This also contributes to the entry costs for establishing a digital library.

It is expected that the entry costs will be reduced when using the DILIGENT infrastructure. In the ideal case the entry is reduced to that of setting up a Virtual Digital Library using the services offered by an existing DILIGENT infrastructures and content and metadata collections that already exist in the DILIGENT infrastructure. Higher entry cost are required, when collections are needed in the VDL that are not yet part of an DILIGENT infrastructure. In this case the integration costs have to be added. However, as part of the DILIGENT project there will also be a development of methods and tools that ease such integration tasks.

### 3.1.2 Case Descriptions

#### DSpace

DSpace (www.dspace.org) is digital library software for managing research materials and scholarly publications. It has been jointly developed by the Massachusetts Institute of Technology (MIT) Libraries and Hewlett-Packard Labs, and is distributed as open source.

The logical information model of DSpace repositories is hierarchical: Communities, such as schools, departments, labs, and research centres, are at the highest level of the hierarchy. Communities may set up several sub-communities and/or collections. The leaves of the hierarchy consist of so-called items that aggregate grouped, related content and associated descriptions (metadata). Multiple media types are supported, although search appears to be restricted to textual fields, and viewing may require third-party application programs.

Functionally, DSpace mainly supports (web-based) submission of items, indexing of, search for, and browsing by Dublin-Core metadata (author, title, date, keywords), and subscriptions to collections. Items can be protected by means of basic access control and DRM mechanisms.

DSpace is fairly well adopted by its targeted market (research institutions and universities); currently there exist reportedly almost 80 Digital Libraries based on DSpace. All these repositories appear to be isolated from each other, i.e., there is no support for search across installations or the creation of virtual collections.

#### Greenstone

Greenstone 3 (www.greenstone.org) is an open-source digital library tool suite developed by the New Zealand Digital Library Group at the University of Waikoto, New Zealand. It aims to empower users, in particular universities, libraries, and other public service institutions to build their own Digital Libraries.

The basic functionality for the end-user with respect to searching and browsing is similar, though it appears to be more configurable with respect to searchable fields. However, Greenstone in addition supports librarians in creating collections from existing collections (including import from and export to DSpace collections).

Like DSpace, Greenstone is fairly well adopted by its targeted market; currently there exist reportedly about 40 Digital Libraries based on Greenstone. Unlike with DSpace Digital Libraries, most of the Digital Libraries based on Greenstone provide a highly customized search and browse interface.

#### OpenDLib

OpenDLib (www.opendlib.com) can handle a wide variety of document types with different format, media, language and structure. The documents of an OpenDLib library are organized in a set of virtual collections, each characterized by its own access policy. Authorized people can define new collections dynamically by specifying definition criteria. In the same digital library, for example, it is possible to maintain a collection of literature accessible to all users and a collection of historical images accessible only to a specific group of researchers. Each collection is automatically updated each time a new document matching the definition criteria is published in the library.

The basic release of OpenDLib provides services to support the submission, description, indexing, search, browsing, retrieval, access, preservation and visualization of documents.

The search service offers different search options. Documents retrieved can be navigated across all their editions, versions, structure, metadata and formats. All the above services can be customized according to several dimensions such as, for example, metadata formats, controlled vocabularies, and browsable fields. OpenDLib also provides other digital library specific functionality, such as the control of access policies on documents, and the management of "user-shelves" able to maintain document versions, result-sets, session results, and other information, etc. In addition, a number of administration functions are also given to support preservation of documents, document reviewing process, introduction of new collections, and handling of users and user group profiles.

#### Fedora

Fedora (Flexible Extensible Digital Object and Repository Architecture, www.fedora.info) is a general purpose digital repository service developed at Cornell University's Digital Library Research Group. It is available as open source.

The logical information model of Fedora is object-oriented. Each digital object has a persistent id, a set of properties, content (one or more data streams), and so-called disseminators, which associate services to view the object's properties and content. In addition, objects can be related to each other to represent, for example, collections and part hierarchies, object derivation (versioning) and equivalence, and data/metadata associations. Some properties (e.g., for realizing authorization policies), and relationships are predefined with special purpose implementation support, others may be user defined. An XML serialization of the object-model (FOXML) making use of RDF, where appropriate, exists.

The functionality of Fedora is realized by means of loosely coupled services. The three main APIs are the Management API devoted to creating and maintaining digital repositories, the Search and Resource Search Index API devoted to simple field search and RDF-based metadata search, and the Access API devoted to view objects and their properties.

Fedora currently reports to be deployed in about 30 digital library projects, which appear to be at different degrees of maturity. The main reason for these projects to adopt Fedora as opposed to (e.g.) DSpace or Greenstone, appears to be the service oriented architecture of Fedora, and the more elaborated object model.

#### Daffodil

Daffodil (Distributed Agents for User-Friendly Access Of Digital Libraries, www.daffodil.de) has been developed at the University of Dortmund, Germany. It is realized as a remote service that is now hosted at the University of Duisburg, Germany, and can be accessed via a client implemented in Java.

Daffodil wraps several existing bibliographic databases (about 20), by forwarding Daffodil queries to these databases and integrating their results. Its logical information model is designed along the lines of the information provided by these sources (authors/editors, documents, classification hierarchies, etc.)

Functionally, the major objective of Daffodil is to support and evaluate advanced search beyond simple fielded keyword search, including so-called search tactics (broadening/ narrowing queries), and search stratagems (e.g., search along citations, and co-author relationships). To this end, the Daffodil client supports a variety of special purpose search and browsing tools, which partially accept input from each other by drag and drop. In addition, Daffodil supports the creation and maintenance of personalized collections.

Daffodil reports to currently have about 100 registered users, and in addition supports temporary accounts for visitors.

#### ALEXANDRIA Digital Library

The Alexandria Digital Library (ADL, http://www.alexandria.ucsb.edu/) is a project at the University of California, Santa Barbara. It offers both, a running service for searching for geo-referenced material, as well as open source software to create and offer geo-referenced Digital Libraries.

The major goal of the ADL is to provide a mapping between geographic names and their geographical coordinates. Consequently, the main distinguishing point of this digital library system is that all items are geo-referenced and time stamped. On this basis, containment relationships between items are derived. Other aspects of the logical information model are similar to that of other digital library systems.

There exist several web interfaces to ADL, providing for combinations of geo-referenced search (by regions) and metadata search. The underlying system architecture is distributed, i.e., collections may be physically distributed. A light weight protocol supported by a java client also allows using ADL as a service.

ADL currently contain close to 4.5 Million geo-referenced items. No usage statistics are available.

### Bricks

BRICKS is an Integrated Project that was selected as one of the projects in the European Cultural field in the 1st call of IST - VI FP. The project aims at integrating the existing digital resources into a common and shared Digital Library, a comprehensive term covering “Digital Museums“, “Digital Archives“ and other kinds of the digital memory systems. The results of the Project will constitute the main assets of a Factory, which has been subsidised by the Consortium partners and the EU for the duration of the Project, but will sustain itself in the future.

The project started on the 1st of January 2004 and will last for four years. It has a budget of 12.2 million Euros, partially supported by EU grant.

The objectives of the project are to design, develop and maintain a user and service-oriented space to share knowledge and resources in the Cultural Heritage domain. The target audience is very broad and heterogeneous and involves cultural heritage and educational institutions, research communities, industries and citizens.

Existing centralised Digital Libraries architectures cannot handle such high level of heterogeneity, so they chose i) a service oriented architecture based on Web services, ii) reuse of existing communication channels and the contents of operational Digital Libraries, and ii) to be as open and flexible as possible, giving the opportunity to join or leave the system at any time without administrative overhead.

With respect to access functionality, BRICKS provides appropriate task-based functionality for indexing/annotation and collaborative activities e.g., for preparing a joint multimedia publication. An automatic annotation service will enable users to request background information, even if items have not been annotated by other users yet. By selecting appropriate items, such as definitions of concepts, survey articles or maps of relevant geographical areas, the service exploits the currently focussed items and the user’s goals as expressed in the user profile. In addition, the linking information, which is generated dynamically, must be integrated into the documents.

### Thematic Realtime Environmental Distributed Data Services (THREDDS)

Just as the World Wide Web and DL technologies have simplified the process of publishing and accessing multimedia documents, THREDDS aims at providing needed infrastructure for publishing and accessing scientific data in a similarly convenient fashion.

Data collections are a cornerstone of environmental research and education studies. New levels of accessing and using data are achievable because of evolving technologies, even as the amount and variety of Earth system data are increasing daily.

THREDDS develop a scientific data web that facilitate the publication, discovery, and use of environmental data, just as the World Wide Web has made the publication of and access to textual and multimedia documents simple and straightforward.

Recent parallel progress in the worlds of scientific data management and education-oriented digital libraries is highlighting a common need to discover widely distributed data sets, and to use unfamiliar data meaningfully with a comprehensive set of analysis tools for:

- Visualizing complex, multidimensional data
- Integrating and overlaying data from multiple sources
- Gracefully handling coordinate systems, measurable quantities, units of measure, and sampling variations. [THREDDS]

### 3.1.3 Services provided

Digital libraries systems may be viewed as an integrated set of software, hardware, and protocol elements that together provide a means of storing, managing and accessing digital documents. All such digital library systems will have some degree of commonality among their functionalities or services; still, they may and do differ considerably in the specifics of their targeted usage scenarios, their emphasis of which features are important, their detailed design choices, and their implementation.

Digital library systems can be categorized along several dimensions, e.g.: proprietary versus open; supporting large scale versus limited collections; supporting many users simultaneously or few, etc. However, the most important dimension concerns what services Digital Libraries offers. A list of the main services in existing digital library systems have been gathered by examining international examples of Digital Libraries. The analysis has been based on web sites examination.

Services have been grouped into categories according to the functionalities provided by Digital Libraries, as follows:

Access & Discovery services

Services providing access to digital cultural resources and disseminate knowledge which can enrich the meaning of research experiences made by users.

- **Access to digital collections**: allows to access digital collections through browser interfaces.

- **Access to catalogue**: allows access to an online catalogue of a collection (i.e., the OPAC) that is available to public, replacing traditional card catalogues.

- **Access to external catalogue/resource**: allows users to access "external" catalogues/resources. Access is managed by grants in a federation of Digital Libraries.

- **Search (simple, advanced) resources**: allows users to search resources (i.e., documents, texts, images, multimedia objects). It may be simple search (i.e., use of keywords) and advanced search (i.e., use of many properties for find a resource).

- **Semantic search of resource**: improves traditional searching enabling computers to work in cooperation through machine readable information.

- **Browse resources**: allows users to discover new information in the Digital Libraries. Users can browse (navigate) among many categories and follow links of results.

- **Cross-search**: allows simultaneous searches across multiple journal databases, book catalogues, and other available information sources.

- **Collection description**: provide users with a description of a collection

- **Download resources**: allows users to download selected resource (i.e., documents, images)

- **Object comparison**: allows users to perform comparisons between different digital objects and view results. Users specify various parameters that form the basis for comparison.

- **Multi-language user interface**: allows users to choose their favourite language for using the application.

Management services

Services that enable activities in the archiving and publishing management processes.

- **Self-archiving**: allows archiving of a digital document in a public accessible website

- **Self-Publishing**: allows the author to have control of the whole creative and selling process (i.e., editing, uploading and publishing)

- **Sharing documents**: allows to share documents directly among network users

- **Metadata creation and management**: offer the possibility to administrate the process of metadata creation, modification and storage. Metadata is "data about data", in other words metadata describes data.

- **Metadata crosswalk:** a functionality that allows to correlate or "map" metadata used in different databases/catalogues allowing search among different Digital Libraries.

- **Print-on-demand**: allows user to print a chosen document/image.

Transactions & Trustiness

Services enable online transactions activities (i.e., e-Commerce) while guaranteeing protection of intellectual property rights (IPR) of digital cultural resources exchanged.

- **Support for e-payment**: allows to manage the e-payment for digital resources online purchasing

- **DRM systems**: helps protect the copyright of materials by defining the policy of content usage

- **Watermarking**: protects distribution of digital resources through a pattern of bits inserted into the digital file (image, audio or video) that identifies the file's copyright information (author, rights, etc.)

- **Authentication**: users, services and resources may need to be authenticated to provide assurance that they are what they purport to be.

### Creation of new content

Services enabling document creation and publication.

- **Annotation**: a note added to a document to provide additional information or comments.

- **Collaborative document creation**: allows different authors to create and modify documents in a collaborative way

### User Registration

Services allowing users to be identified by the system creating a profile for grants and preferences.

- **Basic data**: basic data (mainly personal data) needed to identify the user, its institution or organisation, and reference to contact him (phone number, email, etc);
- **Profile or preferences**: a set of metadata dynamically created allow the dynamic customisation of the interface, improve filtering in search results and push news or other kind information potentially relevant for the user.
- **Grants or permissions**: a set of information required to allow user access resources and services.

### Services for building Community

Services for building Virtual Communities of Digital Libraries users to increase audience and loyalty among users, including Education activities.

- **Chat**: allows participants to communicate by typing messages which are displayed almost instantly on the screens of other participants who are using the chat room

- **Forum**: public meeting or assembly for open discussion

- **Help-desk**: provides service to internal or external callers to facilitate the use of their technology assets

- **Peer-review**: provide with critical evaluations from experts authors of articles prior the publication

- **Recommendation/Rating**: Recommends/rates digital materials, obtained as results from a search activity, in order to rank them. Recommended/rated materials are suggested as favourite results also for other interested users that submit similar queries.

- **e-learning**: provides education services via the Internet using electronic applications and processes for enabling users to learn in an interactively way. Content can be delivered via the Internet, satellite TV, and CD-ROM.

## 3.2 VIRTUAL RESEARCH ORGANISATIONS

While the concept of Virtual Community is widely used and accepted, none referred to Virtual Organisation nor Virtual Research Organisation in any DL project analysed.

One of the characteristics of a virtual group or community is that it is not determined by organizational structures or a common task, but by a common interest or goal. The members of such a community might be distributed across the boarders of organizations and nations, thus don't reveal any form or structure nor hierarchy.

In the DILIGENT context we deal with Research Organisations became virtual by the eInfrastructure. In some way the research purpose is the aim for meeting and the eInfrastructure is the media to meet across any physical barrier.

The creation and operation of such communities can be supported by information and communication services that take into account the special requirements of such communities. After discussing the characteristics of Virtual Communities and Virtual Organizations this section describes example cases that illustrate the different ways Virtual Communities are supported by IT systems. In the last part of the section services are identified classes of services that help in supporting the activities of Virtual Communities.

### 3.2.1 Virtual Communities vs Virtual (Research) Organisations

The phenomenon of Virtual Community (VC) is found far beyond the scientific domain. The first notion of Virtual aggregation or communities is dated to the early 1980s where many people (e.g., computer hackers) found it useful to **meet "on-line"** to exchange information, code, passwords, etc. From this **unorganized form** the VCs have driven demand for the development of more organized services like Bulletin Board Systems (BBS). In this context, the term *PRO-SUMER* has been introduced to characterize the user of a VC that is at the same time producer and consumer of the information exchanged within the community.

The characteristic of a virtual group or community is that it is not defined by organizational structures or a common task, but by a common interest or goal. Several studies in the 90s revealed that the success of some VCs is based on the human needs of aggregation based on **common interests**. Such common interests are the reasons for the creation of the **human network** and the start of a discussion, or exchange of ideas, opinions, documents etc. Since there is no organizational or other formal context for the functioning of a Virtual Community, it is crucial to keep the community members involved and committed to keeping the Virtual Community alive. Active networking, the quality of a discussion, and the amount of information available in the virtual site that can be considered as a representation of the Virtual Community are some of the factors **that keep people** linked with the community and create a virtuous circle. In summary, it is necessary to give people a positive experience of being part of the Virtual Community.

The World Wide Web (WWW) played a central role in enabling Virtual Communities, by providing a platform that makes the provision of information as well as the access of such information by all community members very easy. Businesses also identified a market perspective in the VC, in particular focusing on **the motivation of aggregation**. Any community[1] created around a specific theme is at the same time:

- a set of potential customers,
- a source of feedbacks and reactions which can be used as data to analyse and improve marketing, a product, etc.,
- a starting point for defining and creating a new product.

Besides other areas of human collaboration there are also Virtual Communities in research, which aggregate information for research purposes, are very dynamic in the member composition, and have the dual role of Producer and Consumer of knowledge.

Systems and services supporting Virtual Communities typically provide some type of a common information and communication space, which can be used and updated by the members of the Virtual Community.

A notion of *Organisation* is present in the Enterprises environment evolution neat described by D.J.Skyrme in [SKYRME]

> "As organizations restructure to respond to their environment, there has been a growing recognition of the need for new kinds of organizational structure. **The Networked Organization** is one such response. It has been defined by Lipnack and Stamps as one:
> *"where independent people and groups act as independent nodes, link across boundaries, to work together for a common purpose; it has multiple leaders, lots of voluntary links and interacting levels."*
> Other types of organization have been described, such as the **lattice organization**, the **spider's web**, the **holonic enterprise** and the virtual corporation. All describe new ways of organizing which:
> - gain authority not from a hierarchy but from individual's *recognized* knowledge and skill
> - link people and teams across conventional boundaries (e.g., departments and geographies)
> - have members and structures that adapt to changing circumstances
> - where management is a sense of mutual responsibility vs. following orders
> - explore ways to work effectively vs. following pre-defined processes
> - readjust or disband teams as needed
> - and therefore exhibit characteristics of innovation, resilience, and self-management.
>
> The notion of a network implies **nodes** and **links**. The nodes can be people, teams or even organizations - networks operate at many levels. Common examples are distributed geographic teams in large organizations, or small organizations operating as networks to compete against large corporations. The links are the various coordination and "agreement" mechanisms. In a network, high degrees of informal communications (both face-to-face and

[1] See for instance: the open source community, the BBS communities, etc…

> over electronic networks) achieve success where formal authority and communications in hierarchical organizations often fail. Two way links and reciprocity across the links are what makes networks work"

Skyrme gives evidence to the notion of structure present in any organisation. The structure, imposed or auto-determined is relevant to meet organisation objectives. While in the community any individual still has its own objective that eventually, in a virtuous circle, will allow any other, hence the community, to meet their objectives.

The idea of a **Virtual Organization** in the context of Grid technologies and upon which the idea of Virtual Digital Libraries is built, is defined by the type of resources shared: the shared space that the members operate in is no longer restricted to the exchange of information, but also includes the sharing of resources across organizational boundaries. Indeed, virtual organizations are defined in [Foster+01] as flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources. This still doesn't brings together aspects of real organizations (institutions nor enterprises) with the idea of Virtual Community that act across the boundaries of organizations and other units, because the notion of structure is indirectly included in the notion of institution and the focus is on the resources rather than the members.

In DILIGENT we have identified the three key elements for a Virtual Research Organisation definition:

- *Members*: who form the VRO, hence the community of researchers;
- The *Membership*: the aims in participating at the VRO research objective and the inclusion (in some way) in the hierarchical *structure* of the research team (by competence, by institution, …)

that define the notion of Research Organisation,

- the absence of physical barriers: the eInfrastructure as a virtual meeting point, but even the concept of collaboration across institution and other type of organisations

that define the notion of Virtual

- the type of resources handled: as in the Grid environment members share computing and storage power, but also archives (or content sources) and Digital Library services (as identified in the previous chapter;

that characterise the VRO in the Grid environment.

In the following paragraphs we study some IT solutions supporting VCs in order to identify typical services supplied to final users.

### 3.2.2 Cases Description

Virtual Communities exist in a wide variety of domains and in many different forms. They differ, for example, in the topic the Virtual Community is focused on, in the reach (regional, national, global), in the closeness of the collaboration, in the size of the community, and in the degree of the involvement of the members. These characteristics determine the way such a community sets up and operates its information and communication space in the World Wide Web. A wide variety of systems and services that support Virtual Communities have been developed, reflecting the differences in their requirements.

We decided to select and describe some example use cases that illustrate the capabilities of Virtual Communities and Virtual Community support:

BSCW Server

(http://bscw.fit.fraunhofer.de/index.html):

The BSCW is a system for supporting collaborative work and was developed at the Fraunhofer institute FIT. It is meant for smaller Virtual Communities and teams that work together very closely, although maybe distributed over different organizations and different countries. When a BSCW server is set up it can support one or more Virtual Communities. The BSCW server supports a service for inviting and registering new community members. In the core of the functionality of the server is a structured shared workspace, where members can upload different types of files and make them available to the other members of the community. Furthermore, the BSCW server supports the annotation of documents (adding a note), the set up of discussion forums, the management of community calendars for entering community events, and the management of an address book for making information about community members. In addition, the BSCW server supports awareness services like marking of new unread content and notification about changes in the workspace.

The BSCW system is based on a client-server architecture. The shared information is stored on a centralized server. There are also alternative architectures for such systems. The Groove Workspace (http://www.groove.net), for example, uses a Peer-to-Peer approach for the management of the shared information objects.

Sun Developer Network

(http://developers.sun.com/):

In the area of software development and the use of IT technology many Virtual Communities exist, that operate sites for exchanging information and experiences with the use of the respective technology. An example for such a Virtual Community is the Sun Developer Network where the community members are typically Java Developers. The site of the Sun Developer Network provides a broad information offer including information about new developments, glossaries on relevant technical terms, background information on related topics, newsletters, and FAQ lists. Furthermore, the site provides communication services like discussion forum, and chat sessions on special topics. In addition, the site of the Sun Developer Network also provides more specific information and communication services for this type of Virtual Community members like downloads and bug reporting,

An interesting approach, which was used in the Sun Java Network for some time for increasing community involvement and community based quality control, was the assignment of so-called Duke Dollars. In Java Forum users could pose questions and than assigned an amount of Duke Dollars to the answering experts according to the usefulness of the answer.

The Sun Developer Network is not only an example for a site for Virtual Community consisting of developers, but also for a site, that is mainly operated by one organizations and the community members act as consumers and in some parts also as contributors.

SourceForge

(http://sourceforge.net/):

Source Forge is a system that supports the formation of a Virtual Community around an idea or a development project. The system, which exists as an open source as well as a commercial implementation, enables the definition of SourceForge projects and the collaborative development of such projects by a Virtual Community. The SourceForge site is supporting the Virtual Community of all SourceForge users and also provides an entry point to concrete SourceForge projects, each of which has its own Virtual Communities, i.e., the persons interested in and contributing to this specific SourceForge project. Within a SourceForge project the Virtual Community is supported by different kinds of information and communication services including a collection of important documents, version management for files (software versions), management of bugs (bug reporting and bug tracking), management of feature requests, and discussion forum.

finebrain Expertise Platform

(http://www.finebrain.ch/)

The finebrain Expertise Platform system is a commercial knowledge management system that focuses on the structured organization of information within Virtual Communities. The system arranges for pipelines of information and questions in a network of experts and other community members. The underlying paradigm is the organization of all resources (documents, experts, and community members) based on a taxonomy/partonomy that can be used to structure relevant topics as well as organizational units like work packages of a project.

FoaF project

(http://www.foaf-project.org/)

Whereas most of the other use cases described so far focus on the provision of information and communication services to members of the Virtual Community, this project enables a look on the community itself. It relies on a RDF-based representation of persons using the FoaF vocabulary. This is a vocabulary for describing persons, the people they know, their interests, and other properties. The combination of these RDF based descriptions of different community members (which can be provided by each individual member) enables a variety of applications that use this machine-readable information to make the structure of the Virtual Community visible and to increase awareness for the community.

With the *foafnaut*, an application developed as part of the FoaF project, a community member can create a graphical representation of the network representing the Virtual Community based on the relationship A knows B. The navigation of such a network can provide an improved understanding of the social relationships within a Virtual Community and may provide starting points for coming into contact with community members.

Wikipedia
(http://en.wikipedia.org/):

Wikipedia is an example of a Virtual Community with the goal of collaborative knowledge collection and editing. Wikipedia is a community-authored encyclopaedia which contains about 520.000 articles on various topics structured by taxonomy. The was started in 2001 by Jimmy Wales and Larry Sanger. Articles are written and can be revised by members of the community. Wikipedia is available in several languages (with differing amounts of articles) and runs on the MediaWiki software. In addition to supporting editing of the articles, the Wikipedia site provides further community services like management of users, discussion forum for every article and search for articles.

Virtual community with the goal of collaborative knowledge collection and editing typically do not collaborate too closely. They make use of the complementing competences of the involved community members. In some cases, like in the Open Directory Project (http://dmoz.org/), which uses a Virtual Community to collect important Web pages for a wide range of categories, there are also special roles assigned to some members of the community like, e.g., being the editor/moderator for a category.

### 3.2.3 Services provided

The following groups of services are used to support the formation and operation of Virtual Communities:

Communication Services

Services that support the communication between members of the communities:

- **Chat:** enables the synchronous communication of community members; There can be chat sessions on special topics involving, for example community experts on this topic;

- **Discussion Forum:** enables the asynchronous discussion of topics relevant for the community. Typically, a discussion forum is focused on special topics within the community.

- **Expertise Brokering:** services that mediate between the need for a specific expertise and experts within the community that have the required knowledge or skills; Such services may put the expertise seeker in direct contact to the expert or may forward a query to an adequate expert;

Information Sharing Support

Typically, sites supporting Virtual Communities support an information space that provides relevant information to the members of the community. In many cases, members of the community are also enabled to become themselves content providers sharing information with the other community members. This makes use of the combined expertise of all the members of the communities. This type of community involvement is supported by services like:

- **Support of document upload:** Services that enable community members to upload files into the common information space; different types of editorial and quality control processes may be defined before finally publishing the uploaded files.

- **Version management:** Services for the management of different versions of files in collaborative development of documents, software, etc.

- **Structured shared workspaces**: In addition to simple document upload this service also enables the structuring of workspaces and further operations like deletion, update, annotation, etc.

#### Awareness Services

This group includes services that raise the awareness of community members for situations and developments within the community.

- **Notification services:** Services that actively inform the user about changes in relevant parts of the shared information space, about other relevant changes in the community or about upcoming events; this also includes services for subscribing for the information that is relevant for an individual community member;

- **Supporting community awareness:** Services that make the members aware of the rest of the community, for the structure of the community, and for the role of the individual within the community; this can e.g., be services for visualizing the community;

- **Newsletters and listing of new information:** Services that point members of the community to new information;

#### Management of Events

One of the tasks of a Virtual Community site is to inform community members about events that are relevant for the members of the community. For this purpose, relevant events can be listed or entered into an event calendar by members of the community.

#### Management of users/community members

Some services of Virtual Community sites require a login of community members. Community members may use real names or artificial identities in the community's information space depending on the type of Virtual Community. User identities may be used to provide individual services to community members and to support personal and targeted communication between community members.

#### Community-based Quality Control

When the members of the communities do not only consume, but also provide information for the shared information space, quality control becomes especially important. One option is a quality control by a special group of the community. Alternatively, the quality control can also be taken over by the community itself in some form of self-regulation. This is supported by services like:

- **Annotation Services:** Services that enable community members to provide comments and ratings for information objects in the shared information space; Advanced annotation services enable the annotation of annotations, different types of annotations and annotation links that enable for example the implementation discourses;

- **Collaborative Filtering:** Services that use the ratings of other community members and the similarity between community members as a basis for making recommendations;

#### Services for Improving Community Involvement

Sites that support Virtual Communities heavily depend upon the interest and involvement of the targeted community members. This means the sites have to provide information that is relevant, interesting, and up-to-date. In addition, community members have to be motivated to contribute to discussions, answer queries and chat by adequate services and methods, e.g., providing some form of compensation to contributing community members.

# 4. Analysis of enabling technologies

Several technologies form the starting point of the Digital Library projects and initiatives analysed: SOA, P2P, Web Services, Open Source solutions, etc. In this chapter we identified two different perspectives from which analyse these enabling technologies. The two perspectives are the Architectural point of view and the Functional point of view: in the former we identified the relevant technologies from the architectural point of view, while in the latter we identified type of applications similar to the one supplied by DILIGENT.

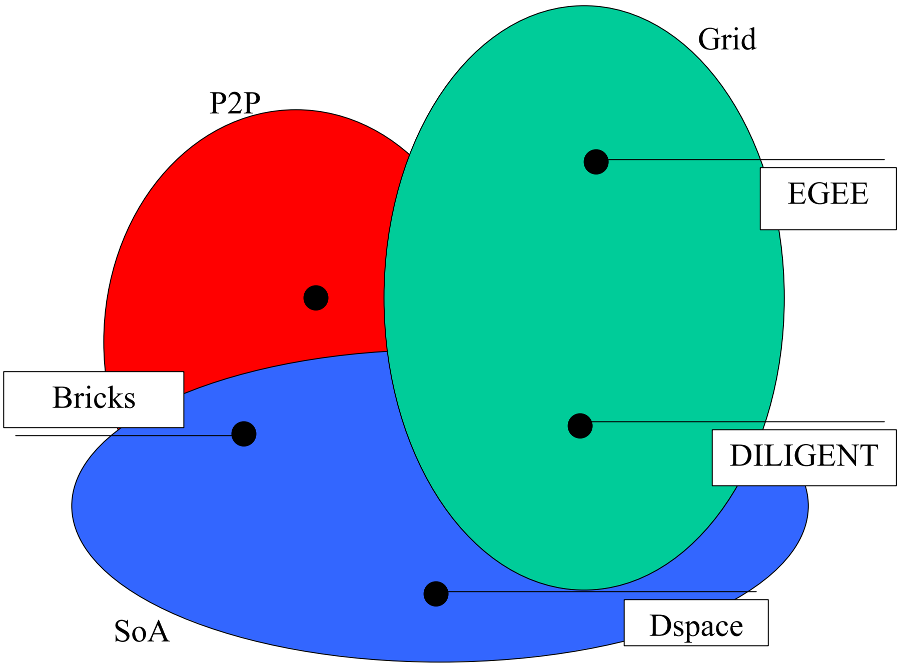
In the following paragraphs we analysed both perspectives and for each perspective we identified the solutions (namely Grid technology and Content Management Systems) that should be considered for the DILIGENT sustainability model.

## 4.1 THE ARCHITECTURAL VS FUNCTIONAL PERSPECTIVE

In any software application the two perspectives collapse in a unicum and often only one is used to qualify the solution itself.
In the field of Digital Libraries the architectures considered can be reduced to be[2]:

- Service-oriented;
- Peer-to-Peer(P2P) enabled;
- Or based on a Grid Infrastructure.

Service orientation is not incompatible with P2P or Grid. As an example, Bricks aims at exploiting a P2P infrastructure to run services, Dspace is based on services but is not fully compliant with WSRF, EGEE aims at creating a Grid infrastructure while DILIGENT propose the creation of a SoA on top of a Grid-based infrastructure (the one created by EGEE).




[2] Heiko Shuldt U.M.I.T. (A) – personal communication

In the following chapters we will describe in detail the Grid technology and its convergence to Web Services.

From the functional perspective we were interested in several Management Systems available for different purposes. From Database Management Systems, to Knowledge Management Systems, passing through Digital Assets Management Systems and others. In particular we have focused our attention on Content Management Systems (CMSs) because in every decomposition of a DL architecture a CMS is explicitly or implicitly (listing the functionalities) referred to.

In this chapter we collect all findings about the analysis of these two technologies (CMS and Grid Computing).

## 4.2 CONTENT MANAGEMENT SYSTEMS

There are no universally accepted standards for what content management systems are or do. Generally, the term "Content Management System" is synonymous with "Web Content Management System", further confusing the consumer. Thus, the definition of a CMS has becomes quite blurred when comparing various vendor solutions. Furthermore, the boundaries of the CMS overlap considerably with document management systems, learning content management systems, knowledge management systems, etc. Still, we know enough to present a strong model for what a content management system should accomplish.

### 4.2.1 Content Management System purpose

Content management as a discipline is the set of processes, technologies, concepts and practices to do with developing, collecting, managing and publishing content. A simple definition of a Content Management System is *a system that manages content*. The main scope of a CMS is to facilitate the publication of contents it manages. CMSs are usually used for producing web-based publications, but they aren't limited to this. For example, a CMS may be used in the case of one print publication. In this case the system converts the needed contents in the format that the print system expects, such as FrameMaker or PDF, which goes to a printer for publication.

### 4.2.2 CMS key concepts

Content

Content is the information that is being distributed. A book, a magazine article, and the information on a web site is all content. Content is information put to use (it can be text, images, graphics, video, sound, etc). Information is put to use when it is packaged and presented (published) for a specific purpose. More often than not, content is not a single 'piece' of information, but a conglomeration of pieces of information put together to form a cohesive whole. A book has content, which is comprised of multiple chapters, paragraphs and sentences. Newspapers contain content: articles, advertisements, indexes and pictures. The newest entry to the media world, the Web, is just the same: sites are made of articles, advertisements indexes and pictures, all organized into a coherent presentation.

Metadata

Metadata is data about data (content). Metadata is the data that we wrap around other data in order to make it more useful and manageable. Typical examples of metadata are: author, insertion date, key words, etc. For one corrected classification of the content it is necessary to have useful metadata and that they describe correctly the content to they associated.

Separation of Content and Presentation

Another fundamental concept of content management is the separation of content and presentation. That is the difference between the 'raw' piece of content itself and the way which it is rendered to whatever medium for consumption. Separating content and presentation makes it easier and more cost effective to manage content. The separation of content and presentation allows for the re-use of content.

Templates

Templates provide a structured way of displaying content. Templates bridge the gap between the neutral content in the repository and the needs of a particular publication. Templates are programs that use either a proprietary or an open programming language to specify how to create the publication through the content available in the system.

Content lifecycle

Content follows a lifecycle. It is created or otherwise sourced and manipulated into the requested form. It then needs to be stored and managed in ways that make it accessible and usable before it is published. Workflow practices occur throughout in order to manage the work of moving the content through its lifecycle as efficiently as possible.

Workflow

Workflow is the series of tasks and triggers that connect the tasks in a particular cycle of work. Workflow is about formalising and codifying work processes. Content management workflows typically contain tasks that change the state of content. That might not be a change to the content itself but a change to the status of the content: approved versus unapproved, for example. The triggers between each task in the workflow can be based on a range of things: dates, content creation or deletion or modification events, system events, or even a manual trigger following a verbal instruction based on a real world event. The notification of change of status, in order to alert the next task owner in the chain, is often done by e-mail but might also be done through another communication system. More generic workflows can be built in a modular way and then combined like building blocks to create more complex workflows.

### 4.2.3 CMS structure

CMS can be considered a software that enables a variety of staff (technical and non-technical) to create, edit, manage and finally publish a variety of content (such as text, graphics, video etc) under the constraint of a centralized set of rules, process and workflows.

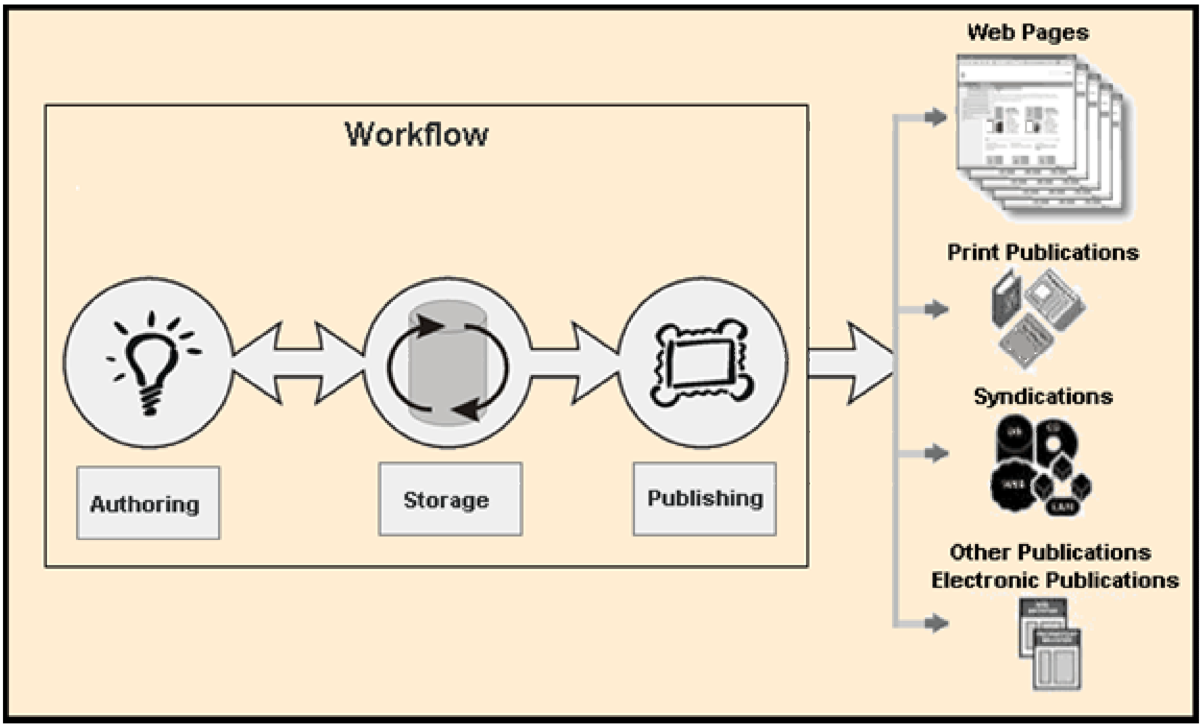
CMS functions can be divided into four categories: **authoring system**, **workflow system**, **storage system** and **publishing system**.

- **The authoring and collecting system** includes the tools and procedures that are employed to gather content and provide editorial and metadata processing. A CMS can help provide an authoring environment as well as input templates (content capture templates) to facilitate the authoring and collecting process.

- **The workflow system** includes the tools and procedures that assure that the entire process of collection, storage, and publication runs effectively, efficiently, and according to well-defined timelines and actions. It permits the definition and management of steps taken by the content in its lifecycle (from authoring to publishing). At each step the workflow system manages the status of the content, notifying (through Communication services) the user involved and escalating jobs where required. In this way, the workflow capabilities allow more authors to be involved in the management of the contents, while maintaining strict control over the quality, accuracy and consistency of the information.

- **The storage system** is the repository of all content and metadata information. The repository can be made by one or more databases; it can include the file system or some network resource. The repository must allow to search and retrieve stored contents. The storage system also keeps track of content versions and who changed what and when (Auditing). The storage system manages versions, allows to recover one previous version of a whichever object (Version control - rollback), and ensures that no two users modify the same document at the same time (lock-in/out).

- **The publishing system** is the process by which stored content is delivered. The publication system is responsible for pulling contents out of the repository and creates publications. Traditionally this has meant 'delivered to the web site as HTML'. However, it could also mean creating an e-mail message, a PDF file, or a Wireless Markup Language (WML) file to be displayed on a mobile phone in a WAP browser.




### 4.2.4 Similarities with other solutions

There are applications that are somewhat related to CMS, but that differ in key areas. A typical CMS is not:

- **A manipulation tool.** A CMS is not a text, graphics, video or audio manipulation tool. It manages the content created by such tools.

- **A transaction processing tool.** CMS do not process transactions in the way an e-commerce engine does. The CMS can manage the product and pricing information as well as any other elements of content the user experiences in the online transaction process, but it does not do the actual transaction processing.

- **Only a publishing tool**. A CMS contains more functionality that publishing only. The role of the CMS in collecting content is often underestimated. Only by collecting content in the correct way can it later be manipulated and published as required.

The identification or name used for a CMS or CMS-like product depends on several factors. These factors include the type of information in the repository, the intended users of that information and the marketing strategy of the CMS vendor. For example basic CMS functionality is found in the following systems: MAMS, ECMS, DMS, IMS, KMS, LCMS. Armed with an application-specific functional name, a vendor can add a trademarked name to further differentiate his product from the competition. Thus, a newcomer to the field may be unaware that the brightly wrapped product is, in fact, a CMS. Some vendors may argue that their products are not content management systems - because of unique features or enhancements. What distinguishes one CMS from another? All CMS' rely upon repositories for storage. Most have workflow capabilities. And most have some built-in search capability. These are obvious similarities. So, how do CMSs differ? The answer is the type of data in their repositories.

Here is a very brief introduction to DMS, MAMS, LCMS and KMS.

**Document Management System (DMS)**: DMS' share many of the characteristics of a CMS in as much as they help store, retrieve and provide access to content (document) but there are two important differences. First, document management systems focus on handling files, of any type, rather than content. Secondly, a content management system is expected to create dynamic publications whereas a document management system is only for managing and retrieving files efficiently.

Document Management Systems can be characterised as follows:

- focused on managing documents, in the traditional sense (like Word files);
- each unit of information (document) is fairly large, and self-contained;
- focused primarily on storage and archiving;
- targeted at storing and presenting documents in their native format;
- limited web publishing engine typically produces one page for each document.

It should be noted that many of today's foremost content management vendors come from a document management heritage.

**Media Asset Management System (MAMS)** : Also known as simply Asset Management System (AMS) or Digital Asset Management System (DAMS), it is especially suited to managing multimedia content and tend to offer hooks into specialized desktop media authoring systems. It places additional emphasis on allowing file manipulation and conversion (for example converting GIF files to JPEG). A 'media asset' consists of multimedia unstructured information objects (like pictures, diagrams, presentations or rich-media contents like audio and video) and of structures meta-information. Only by adding the structured metadata a media information object becomes a "value containing asset", which becomes available, can be investigated, reused, and traded.

**Learning Content Management System (LCMS)**: LCMS is a system (mostly web-based) that is used to author, approve, publish and manage learning content (more specifically referred to as learning objects). A LCMS combines the administrative and management dimensions of a traditional LMS (Learning Management System) with the content creation and personalized assembly dimensions of a CMS. Just like in a CMS, there would be workflow processes around LCMS.

An example:

- Instructional designers would create either new RLOs (Reusable Learning Object) targeting specific performance goals, or new courses by assembling already created RLOs.

- Editors would view the submitted RLO/course and either approve or reject it. If approved the RLO/course would be made available to all to use, otherwise it would be sent back for revision.

- Personalization rules would set in, targeting the new RLOs/courses to those who fit (or have subscribed to) its profile.

- RLOs and courses that have outlived their usefulness would either be archived or deleted from the repository.

**Knowledge Management System (KMS)**: Knowledge management is about trying to understand and harness the power of the knowledge that an organization has. Knowledge management focuses on the tools and practices for discovering and squeezing out an organization's most valuable knowledge assets and making them available to the organization. The tools use technologies to search, index, mine, analyse, categorize and synthesize vast amounts of data into useful knowledge. However, knowledge management is not about authoring and collecting content in the same way as content management. Nor are knowledge management systems usually as advanced in their management and publication capabilities as complete content management systems. Whereas knowledge management focuses on addressing the challenge of identifying and distilling 'knowledge', content management is the infrastructure that makes it possible to store and distribute content. There is not reason why a content management system should not be used to manage and publish the knowledge created by a knowledge management tool.

### 4.2.5 Content Management System core functions

There is not single list of requirements for a content management system. Every organization has unique needs. Each product has its own set of strengths and weaknesses, and distinctive design principles.

The purpose of this chapter is to describe a set of requirements that, in our vision, a system has to provide to be called CMS.

A CMS typically provides these functions:

Authoring and collecting

**Authoring**: This is the process of creating content from scratch.

**Acquisition**: Here you gather content from some existing source.

**Conversion**: In many cases, much of your planned content will be based on existing content that is in an array of formats and structures and needs converting and processing before it can be managed within the CMS. Structure needs to be added to the content, or re-mapped to fit your content model, and unnecessary data discarded.

**Adding metadata**: Either at the point of creating the content, or subsequently as part of an editorial and review process, metadata should be added to the content in order to manage and publish it within the system.

**Aggregation**: the process of choosing and composing contents to create new kind of content

Workflow

**Communication**: It supplies the services that allow the communication between user and user or between system and user. It can moreover allow the integration of other channels of communication (such as e-mails).

**Workflow management**: It permits the definition and management of steps taken by the content in its life-cycle. At each step the CMS manages the status of the content, notifying (through Communication services) the user involved and escalating jobs where required. In this way, the workflow capabilities allow more authors to be involved in the management of the site, while maintaining strict control over the quality, accuracy and consistency of the information.

Storage

**Store content**: The repository may be one or a set of databases of various kinds. It can include the file system or some network resource. The repository must be able to store: textual content, file-based data and meta-information.

**Versioning and rollback**: It keeps track of all the versions of a content, and who changed what and when (Auditing). Manages the versions, allows to recover one previous version of a whichever object (Version control - rollback). Ensure that no two users modify the same document at the same time (lock-in/out).

**Classification, Indexing**: CMS often provide support for integrated indexing and classification functionality.

**Search and retrieve**: With all the content that the CMS manages, you need effective mechanisms for finding what you need.

**System configuration and administration**: The CMS itself needs configuring and administering.

**Logging**: Keeps track of all the operations made within the system, who made what, and when.

**Backup**: Provides backup services.

Security

**User and Group management**: Provides the services to define and manage the users who can access the system. Users can be organized in groups. It also permits to organize users in groups.

**Permission and role management**: Define and manage the CMS privileges (what operations are allowed to perform) that can be assigned to users or groups.

**Authentication**: The process of determining that a user is who he/she claims to be. Users are typically authenticated by a username and password, or similar mechanism.

**Session management**: Preserve a client "state" from one session to the next, usually through cookies.

**Cryptography**: Coding of private information (generally user data).

Publishing

**Deployment**: Deployment refers to the processes and practices involved in moving content from a development environment, such as a development or staging server, to the live environment. Deployment seeks to ensure that the right content is transferred at the right time. In some cases, deployment is not a defined process step (example: as soon as content goes live on the system it also goes live on the site). In other cases a CMS may be used to publish full versions of the site which are first reviewed and tested in a development environment before being deployed to the live servers according to a release process.

**Accessibility**: The CMS must conform to standards such as the W3C Web Accessibility Initiative (WAI).

**Cross browser support**: The pages must be viewable in all major web browsers (Internet Explorer, Netscape, Opera, etc).

**Stylesheets**: Final appearance is controlled through the use of stylesheets. This provides flexibility and expandability.

**Page templates management**: Overall page layout is specified via page templates. This function permits to define page templates. Ideally, a non-technical interface should be provided for managing this. Page generated by templates must include stylesheets described above.

**Personalisation**: Different information is presented to different users based on their user profiles.

### 4.2.6 Content Management System vendors

After we have seen the main characteristics that characterises a CMS, it is useful to see the main products available in the market.

Commercial solutions

Below are some of the leading CMS vendors broken down into three broad categories based on cost and solution completeness. Enterprise solutions typically contain features and capabilities that mean they are not considered only CMSs but e-business platforms that incorporate CRM, CMS, measurement and more.

- Enterprise scale solutions
  - Vignette - http://www.vignette.com
  - Broadvision - http://www.broadvision.com
  - Documentum - http://www.documentum.com

  - Interwoven - http://www.interwoven.com
  - Divine - http://www.divine.com
- Mid-market solutions
  - Microsoft - Content Management Server - http://www.microsoft.com/cmserver
  - Obtree - http://www.obtree.com
  - Mediasurface - http://www.mediasurface.com
  - Stellent - http://www.stellent.com
  - Tridion - http://www.tridion.com
  - Reddot - http://www.reddot.com
- Lower price solutions
  - Infosquare - http://www.infosquare.com
  - Atomz - http://www.atomz.com
  - UserLand - http://www.userland.com
  - iBlurbs - http://www.iblurbs.com

Open source solutions

Some of the most promising open source projects are the following:

**TikiWiki**

Tiki software (aka TikiWiki) is defined by its developers as an *"open-source Content Management System (CMS) and Groupware that can be used to create all sorts of Web applications, Sites, Portals, Intranets and Extranets"*

TikiWiki works great as a Web-based collaboration tool. It's a multi-purpose package with a lot of native options and sections that you can enable or disable as you need them. It is designed to be international, clean and extensible.

Major features include articles, forums, newsletters, blogs, a file/image gallery, Wiki pages, drawing applet, trackers, a directory, polls/surveys and quizzes, a FAQ, chat, a banner management system, Webmail, a calendar, Ephemerides, maps, charts, Mobile Tiki (PDA and WAP access), RSS feeds, a category system, a theme control centre, workflow, live support, Shoutbox, ACLs, and more.

**PHPnuke**

PHPnuke is an open source solution written in PHP, which provides many tools to create a site/portal of content.

PHPnuke's web interface has been developed following the suggestions of the W3C.

To be able to use PHPnuke you need an SQL database server (highly tested on MySQL), Apache web server and PHP version 4.x installed and running. From version 5.3 PHPnuke supports MySQL, mSQL, PostgreSQL, ODBC, Sybase and Interbase servers.

The functionalities that can be found in the standard version of PHPnuke are many, including:

- Portal administration through web interface
- Contents insertion by managers or single users, also not registered. Possibility of moderation, approval, modification of the contents
- Possibility to insert comments and to vote every content
- Internal search engine
- Registered users management
- Users area in which every registered user can manage his e-mail, the private messages and the personal configuration of the site
- Customizable user interface
- Download area with 3 protection levels : all users, only registered users and administrators users
- Forum with internal search engine
- FAQ management
- Statistics management
- Possibility to create, manage and estimate polls
- Support of the RSS/RDF format for the export of the news so as to make them available from other portals.
- It is possible to add new functionalities through addition of appropriate modules. Also these modules are open source and freely usable.

**Zope**

Zope is an *object oriented* development platform for web applications. It's written and extensible using Python (scripting language). Zope runs on UNIX platforms as well as Windows NT and is composed of components which work together to provide a complete web architecture:

- **ZServer** - Zope contains an internal web server able to support all the functionalities for developing applications.

- **Web Server** - If necessary it's possible to substitute the internal web server with another one (such Apache) supporting CGI interface.

- **Zope Core** - This is the mechanism everything coordinates, managing the web interface and the objects database.

- **ZClasses** - Zope allows the site administrators to add new objects types through the web interface. These objects are named ZClasses.

- **Zope Products** - Zope Products extend the Zope core by adding new object types and custom facilities written in Python.

- **Object Database (ZODB)** - Internal object database.

- **Relational Database** - The data needed by applications can be stored in the internal object database. Also, it is possible to replace the internal object database with any DBMS for which an ODBC driver is available.

Zope provides one user interface to manage and develop applications. This interface can be used by any modern browser and is very familiar to the users, showing the hierarchy of the objects through one tree structure, in a similar way to like the most diffuse file manager.

One of the most powerful aspects of Zope is known as "Acquisition". The concept of acquisition works with all Zope objects, and provides a powerful way to centralize common resources. The core concept is simply that:

- Zope objects are contained inside other objects (such as Folders).

- Objects can "acquire" properties, content and behaviour from their containers.

## 4.3 GRID TECHNOLOGY TODAY

### 4.3.1 Grid effort in EU

The presence of commercial European entities in the Grid product market is almost non-existent. All Grid computing middleware are the results of research projects based on the North-American Globus Toolkit 2, such as Unicore, LCG2, among others.

Relevant standards are defined below in Section 4.3.5.

The EGEE project

In April 2004, a promising European project started with the aim of creating a production Grid (within the eInfrastructure) - the EGEE (Enabling Grid for E-sciencE). EGEE will integrate several best practices from other EU research projects and provide an open-source Grid middleware (called gLite) for the construction of eInfrastructure. EGEE is a 100 million euro initiative co-funded by the European Infrastructure EC for Research (31,8 million euro). It involves more than 70 universities and industries, and is coordinated by CERN. Some partners are from Japan, Russia, and USA, giving a global scope to the initiative.

EGEE's middleware is called gLite and it "provides a cutting-edge, best-of-breed framework for building Grid applications tapping into the power of distributed computing and storage resources across the Internet." [gLite]. The gLite middleware is a Service Oriented Grid middleware providing services for managing distributed computing and storage resources and the required security, auditing and information services. It is based on some major guidelines such as quick and easy deployment of a set of services, to be as interoperable and fault tolerant as possible, to use existing services whenever possible, etc.

The gLite system is composed of a number of high level services: computing element, data management, accounting, logging and bookkeeping, information and monitoring, security and workload management. These services can be installed on individual dedicated computers (nodes) or combined in various ways to satisfy site requirements.

It is indeed the most interesting project in Europe related to the implementation of a production Grid middleware, but more interesting is the aggregation of computing and storage resources, concretely available for running the Grid in a production environment. More than 9000 CPUs are integrated in the EGEE infrastructure and run LCG2. All nodes will shortly upgrade to gLite 1.0. Three communities (about 3300 users) are using, hence heavily testing, both the infrastructure and software.

The first release of gLite is planned for 2$^{nd}$ quarter of 2005 but will not be Grid Service Architecture (OGSA) compliant. In the next months we will know the exact plans of gLite development team with regard to OGSA and WSRF. Indeed the gLite members are active within the GGF[3] and they push on standardization (Fabrizio Gagliardi - EGEE project manager - is involved in the GGF advisory board).

The EGEE project has established relationships and is working closely with other Grid initiatives and research networks. The EGEE infrastructure builds on the EU Research Network GEANT and exploits Grid expertise that has been generated by projects such as the EU DataGrid project, other EU supported Grid projects and the national Grid initiatives such as UK e-Science, INFN Grid, NorduGrid and US Trillium. The infrastructure aims at providing interoperability with other Grids around the globe, including the US and Asia, contributing to efforts to establish a worldwide Grid infrastructure.

Industry is also involved through the work carried out by the Industry Forum. The main role of the Industry Forum in the EGEE project is to raise awareness of the project amongst

---

GGF is a community-initiated forum of thousands of individuals from industry and research leading the global standardization effort for Grid computing . See paragraph 3.3.3[3] industry and to encourage businesses to participate in the project. This will be achieved by making direct contact with industry, liaising between the project partners and industry in order to ensure businesses get what they need in terms of standards and functionality and ensuring that the project benefits from their practical experience. Industry will benefit significantly from EGEE in three ways:

- Industry as a partner - through collaboration with individual EGEE partners, industry has the opportunity to participate in specific activities, thereby increasing know-how on Grid technologies.

- Industry as a user - as part of the networking activities, specific industrial sectors will be targeted as potential users of the installed Grid infrastructure, for R&D applications.

- Industry as a provider - building a production quality Grid will require industry involvement for long-term maintenance of established Grid services, such as call centres, support centres and computing resource provider centres.

The EGEE project is also making significant contributions to standards bodies such as the Global Grid Forum, and to policy developments in areas such as Grid security, resource scheduling and certificate authorities, which will have an impact on the development of industrial Grid standards.

Also, the EGEE project is likely to be a source of innovative IT technologies. This spin-off is anticipated to have benefits for industry and commerce going well beyond scientific computing, in much the same way that the World Wide Web, initially conceived for science, has had a much broader impact on society.

Regarding collaboration with the user communities, EGEE has set up a strategy to help integrate new user communities on the Grid infrastructure. This strategy consists on allowing new scientific communities to get integrated into the project networking activities bringing new requirements. These requirements are then taken into account by the project middleware activities in defining future developments and priorities. Networking activities include training of the user communities and dissemination of results, which are used to attract new scientific communities.

A large number of Grid research projects have started in Europe. Their major objective is to virtualize resources empowering individuals and organisations to create, provide access to and use a variety of services, anywhere, anytime, in a transparent and cost-effective way. This will realise the vision of a knowledge-based and ubiquitous utility. Between them, EGEE certainly represents the larger collaboration bridging together many distributed resources to build an extended and powerful infrastructure. Many of these Grid projects are presented here: http://entertheGrid.com/

LHC

The mission of the LHC Computing Project (LCG) is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the Large Hadron Collider.

The Grid middleware used for the LCG is based on EGEE developments. At present the EGEE/LCG production infrastructure runs LCG2 Grid middleware and includes components that have been developed by a number of projects and organisations:

- The Globus Toolkit (GT2) developed by the Globus Project
- The Condor system developed at the University of Wisconsin, Madison
- The Globus and Condor components and some other tools from US projects are integrated and packaged as the Virtual Data Toolkit by the VDT project at the University of Wisconsin, Madison. VDT provides support for this package to LCG/EGEE.
- Tools developed by the DataGrid Project (EDG). The EU-funded DataGrid project ended in 2004, but the institutes that had developed the tools needed for the LCG/EGEE Grid continue to support them until they are replaced by improved software.
- New middleware components developed as part of the gLite toolkit by the EGEE project.

#### SEE-GRID

SEE-GRID intends to provide specific support actions to pave the way towards the participation of the South-Eastern European countries to the Pan-European and worldwide Grid initiatives. This is accomplished through dissemination conferences and training material including cookbooks, pilot and demonstration test-beds for hands-on experience, applications' adaptations to be able to use the Grid, operational and support centre schemes and organisation, and finally feasibility studies and roadmaps for the integration of the South-Eastern Europe to the European Research Area (ERA) via an extended Pan-European Grid infrastructure.

SEE-GRID deploys EGEE production middleware (LCG2) and develops two new applications: a Search Engine for South-Eastern Europe (SE4SEE) and the Volumetric Image Visualization Environment (VIVE).

#### GRACE

GRACE is an innovative Information Retrieval system that enables just in time, flexible allocation of data and computational resources, thanks to integration with GRID technology. GRACE aims to make terabytes of information distributed on vast amounts of geographically distant locations highly accessible.

GRACE Application is a decentralized search and categorization engine built on top of GRID technology. The resulting index will allow querying on demand from any other node in the GRID. GRACE is based on LCG2 middleware and deployed in the EGEE dissemination infrastructure (GILDA). Integration of GRACE functionalities in Grid middleware could be foreseen for the future.

### CROSSGRID

The CrossGrid project is oriented towards compute and data-intensive applications that are characterized by the interaction with a person in a processing loop. Such applications require a response from the Grid to an action by a human agent in different time scales. There are many large-scale problems which require new approaches to computing, such as earth observation, environmental management, biomedicine, industrial and scientific modeling. The CrossGrid project addresses realistic problems in medicine, environmental protection, flood prediction, and physics analysis. These applications are oriented towards specific end-users.

CROSSGRID is based on LCG2 middleware provided by the EGEE project. CROSSGRID also developed the following components:

- OCM-G: stands for On-line Monitoring Interface Specification and is used to express monitoring requests. Common protocol and design of monitoring system allow using multiple monitoring tools altogether providing interoperability between them.

- CROSSGRID SCHUDULER: supports a Grid infrastructure for complex applications that are computationally intensive and written with the MPI library.

- MIGRATING DESKTOP: is a framework, graphical user interface for application management, Grid and job monitoring, data and metadata management. By the Roaming Access Server, it includes an authentication mechanism and advanced Grid tools as well.

### OGSA-DAI

The OGSA-DAI project is concerned with constructing middleware to assist with access and integration of data from separate data sources via the Grid. It is engaged in identifying the requirements, designing solutions and delivering software that will meet this purpose.

OGSA-DAI OGSI Release 5.0 (based on Globus Toolkit 3.2). Provides an extensible framework that allows access to and updating of data resources. Inherent data resource capabilities may be exploited and exposed to the Grid, such as performing SQL queries or XPath statements on data, as well as providing additional data manipulation functionality, such as asynchronous delivery or data transformation, that operates at the service layer. Using the OGSA-DAI interfaces disparate, heterogeneous data resources may be accessed in a uniform manner. It is important to note though that the underlying data model is not hidden - a relational data resource remains a relational resource and will not masquerade as an XML data resource unless that mode of access is already supported by that data resource nor does OGSA-DAI provide a query language that is not already supported by the data resource or through some other third party application.

### MAMMOGRID

The aim of this project is to develop a European-wide database of mammograms. This database will be used to investigate a set of important healthcare applications and the potential of the GRID to support effective co-working between healthcare professionals throughout the EU.

MAMMOGRID application is based on AliEn Grid middleware (developed at CERN).

### GRIDLAB

GRIDLAB produces a set of application-oriented Grid services and toolkits providing capabilities such as dynamic resource brokering, monitoring, data management, security, information, adaptive services and more.

GRIDLAB is based on Globus middleware provided by the Globus Alliance. GRIDLAB also developed the following components:

- GAT: services are accessed using the Grid Application Toolkit (GAT). The GAT provides applications with access to various GridLab services, resources, specific libraries, tools, etc. in a way that the end-users and especially application developers can build and run applications on the Grid without needing to know details about the runtime environment in advance.

- GRIDSPHERE: is a portal framework provides an open-source portlet based Web portal.

- GRMS: is a job scheduling and resource management framework, which allows users to build and deploy job and resource management systems for Grids.

- iGRID: is a novel Grid Information Service. Among iGrid requirements there are performance, scalability, security, decentralized control, support for dynamic data and the possibility to handle user supplied information.

- MERCURY: provides a general and extensible Grid monitoring infrastructure. Mercury Monitor is designed to satisfy specific requirements of Grid performance monitoring it provides monitoring data represented as metrics via both pull and push model data access semantics and also supports steering by controls.

- DATA ACCESS AND MANAGEMENT: provides the infrastructure to efficiently manage and transfer large data sets in the Grid. The core of GridLab Data Management infrastructure is the following set of services: the Replica Catalogue for tracking and organizing the Grid filespace, File Service for providing efficient data transfer and StorageBox for data archival and registry.

### UNIGRIDS

UNIGRIDS is an EU funded 6th framework Specific Targeted Research Project (STREP), funded for 2 years from July 2004. UniGridS is a follow-on project from GRIP (Grid Interoperability Project) which was funded for 2 years by the EU to realise the interoperability of Globus and UNICORE and to work towards standards for interoperability in the Global Grid Forum.

The UniGrids project will develop a Grid Service infrastructure compliant with the Open Grid Service Architecture (OGSA). It is based on the UNICORE Grid software initially developed in the German UNICORE and UNICORE Plus projects.

UNICORE provides the users with a rich set of functions to create and manage complex batch jobs that can be executed on different systems at different sites. The UNICORE software takes care of the necessary mapping of user request to system specific actions. Transfer of data between systems and sites is performed automatically by UNICORE. UNICORE ensures that only properly authenticated and authorized users may access resources.

### 4.3.2 Grid solutions in USA

Ian Foster, Carl Kesselman and Steve Tuecke where the inspirers of Grid computing with their papers: "the anatomy of Grid" and the "physiology of Grid". The first two, after a decade working on parallelism promoted the Globus project with the aim of develop a Grid toolkit for Grid enable applications.

Nowadays the Globus toolkit (in 1st quarter of 2005 it is expected to be released an OGSA compliant version 4) is the base product for USA companies selling their own Grid middleware or Grid-enabled application.

The Globus toolkit is developed by the Globus Alliance**,** a consortium dedicated to collaborative design, development, testing and support of the open source Globus Toolkit. The Globus Alliance was established in 2003 by the U.S. Argonne National Laboratory, the University of Southern California's Information Sciences Institute, the University of Chicago, the University of Edinburgh and the Swedish Center for Parallel Computers (PDC). Globus aim is to create the fundamental technologies needed to build computational Grids.

Recently (February 2005), Hewlett-Packard, IBM, Intel and Sun Microsystems announced the formation of the Globus Consortium, a new industry group dedicated to the commercial advancement of the Globus Toolkit, an open de-facto standards building block for enterprise-level Grid implementations.

While the Globus Consortium is not a standards body, it aims to work in tandem with existing Grid standards bodies -- such as the Global Grid Forum (GGF) -- to further drive implementation of standards that best support enterprise Grid implementations.

The Globus Consortium, with funding and engineering support from its member companies, will focus on providing resources and direction on the technology roadmap for the Globus Toolkit. It will define specifications and requirements for the Globus Toolkit, fund code contributions, and deliver leadership on open standards for industry Grid standards organizations.

Another industry-based consortium is the Enterprise Grid Alliance, created in 2004, a consortium formed to develop enterprise Grid solutions and accelerate the deployment of Grid computing in enterprises. The aim of this consortium is to promote a de-facto standard within the enterprise where applications are free from other limitations (mainly security and accounting) when approaching the inter-enterprise market.

All these companies promote the basic USA middleware that is Globus Toolkit, developed by the Globus Alliance, with different packaging and consultancy service for specific application development.

With some differences we should highlight three companies and their initiatives in the Grid world: a Grid-based database and application server (the Oracle 10g) that seems to have only a touch of Grid expected functionalities, a Grid-based services for computing and storage on-demand (SUN Grid), that consider a Grid-based infrastructure completely external to customer ownership, and a network centric operating system – following rumours on the net – that will bring the giant Microsoft on the Grid edge. They are interesting because they demonstrate the possibility for commercialising a product and a service even with the above mentioned limitation in technology and concrete marketing initiatives quite far from the expected Grid solutions in the research field.

### Oracle 10g

Oracle Corp.'s Application Server 10g is easily one of the biggest, most ambitious and most complex e-business platforms, and stands out even when compared with IBM's WebSphere.

It's not 100 percent correct to refer to Application Server 10g as an "application server", even if it offers powerful Java server capabilities that leverage the Grid features of the 10g platform, providing excellent scalability. It has very good standards support in the areas of Web services and Java development—the product provides excellent support for J2EE (Java 2 Platform, Enterprise Edition) 1.4, has some nice features for application management through Java Management Extensions and supports several cutting-edge Java Specification Requests.

And with a typical implementation of the enterprise version starting at approximately $20,000, Application Server 10g is a good value, especially if one takes into account how much it would cost to purchase the aforementioned competing applications separately.

Once Application Server 10g has been effectively and fully deployed, its capabilities are tightly integrated. However, getting to that point entails quite a bit of work. Some users say Application Server 10g may be overkill for the average company. Although Oracle has made great strides since the 9i era in easing the deployment of its enterprise applications, installing and setting up an Application Server 10g infrastructure was still more difficult, time-consuming and frustrating than setting up a comparable IBM WebSphere or BEA Systems Inc. WebLogic infrastructure.

Oracle's vision of Grid capabilities are essentially the scalability and fault tolerance given by hiding the physical management of DBs. This vision is not at all similar to initial ideas by Foster and Kesselman, nor is in line with evolution proposed by GGF or EGA about the convergence to Web-services. The aggressive marketing of the Grid label by Oracle on its products anticipates the market and creates a customer demand for "Grid-based" solution without any clear idea of what "Grid" really means.

### SUN Grid

When evaluating Sun's new service model, do not be misled by the "Sun is proposing its Grid" label. Gartner defines the Grid engine SUN N1 as "a collection of computer resources, owned by multiple organizations, that is coordinated to allow them to solve a common problem."

The Sun Grid is utility or pay-per-use computing, not customer ownership of computing systems. The Sun Grid consists of Sun hardware and software, including its servers based on Advanced Micro Devices' Opteron chip servers, StorEdge 6920 storage arrays and 10 GB Ethernet-networks which run Solaris 10, Java and Sun's N1 real-time infrastructure technology. Customers purchase the CPU hours and storage they require at Sun's Web site. At slightly more than the 80 cents per gigabyte per month Sun charges to have its 6920 system deployed at a customer site, Storage Utility sounds compelling. It costs four to five times less than what managed hosting centres typically charge for shared storage pricing.

Other considerations include:

- The applications customers execute on the Sun Grid must be based on Sun specifications.
- Troubleshooting and monitoring workloads is more difficult in such a fully virtualized and dynamic environment.
- Most per-CPU software licensing contracts could easily negate any potential savings. (To mitigate this, Sun encourages developers to write applications that will become part of the stack on Sun's utility systems.)

### 4.3.3 Grid initiatives in Japan

Japanese industries and universities are collectively making an effort to promote a Grid-based infrastructure for e-Science. The name of the project is NAREGI: "National Research Grid Initiative. The Center for Grid Research and Development conducts research and development of the Grid software infrastructure and network technology that will become the cornerstone of information technology in the 21st century.

Initially the project will exploit existing technology, namely Unicore (Fujitsu), Globus and Condor (University of Wisconsin), but the objective is to develop a new Grid middleware to overcome the actual limitations of computational Grid middleware.

Japanese industries involved in the project don't have specific products but use Globus 2.2 and act as system integrators on specific customer solutions. This is the case for Fujitsu, NEC and Hitachi.

### 4.3.4 Grid effort in the emerging markets

The Chinese Government has assigned great importance to Grid technologies: A special Grid project has been included in the National 863 Program, the China National Grid has been established, the Ministry of Education has set up ChinaGrid special project, and the Shanghai Municipal Government has started ShanghaiGrid Plan for fully supporting research and application of Grid.

India is entering the Grid computing competition through alliances with USA Grid providers (i.e., Satyana Ltd and United Devices) or branches of multinationals (i.e., IBM India, Oracle India, and SUN).

### 4.3.5 Grid Standardization Effort

The Grid vision requires protocols that are not only open and general-purpose but also *standard*. It is standardization that allows potential collaborators to establish resource sharing arrangements quickly and easily with any interested party. It is standardization that allows organizations to establish resource sharing contracts routinely for acquiring resources on demand, thus avoiding the need to build expensive data centres designed to handle peak loads but that remain underutilized most of the time. Standard solutions can allow us to move away from today's often incompatible and non-interoperable distributed systems, toward a model where computing and data capabilities are available as standardized, interchangeable commodities.

Several groups, consortiums, communities and forums are discussing the directions to be followed as part of the standardization effort that is leading the future of Grid computing. Of all consortiums currently working in Grid standards, some are already established groups which are now moving their efforts in this new direction, while others are recently formed groups, fully Grid oriented. The most relevant and well known are:

Global Grid Forum (GGF) [GGF]

GGF is a community-initiated forum of thousands of individuals from industry and research leading the global standardization effort for Grid computing. GGF's primary objectives are to promote and support the development, deployment, and implementation of Grid technologies and applications via the creation and documentation of "best practices" - technical specifications, user experiences, and implementation guidelines. GGF efforts are also aimed at the development of a broadly based Integrated Grid Architecture that can serve to guide the research, development, and deployment activities of the emerging Grid communities. Defining such architecture will advance the Grid agenda through the broad deployment and adoption of fundamental basic services and by sharing code among different applications with common requirements. Wide-area distributed computing, or "Grid" technologies, provide the foundation to a number of large-scale efforts utilizing the global Internet to build distributed computing and communications infrastructures. As common Grid services and interoperable components emerge, the difficulty in undertaking these large-scale efforts will be greatly reduced and, as importantly, the resulting systems will better support interoperation.

Organization for the Advancement of Structured Information Standards (OASIS) [OASIS]

OASIS is a non profit international consortium that drives the development, convergence, and adoption of e-business standards. The consortium produces more Web services standards than any other organization along with standards for security, e-business, and standardization efforts in the public sector and for application-specific markets. OASIS was founded in 1993, in these latest years has started working in Grid computing for the definition of components as part of the WSRF framework.

Internet Engineering Task Force (IETF) [IETF]

The IETF is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. It is open to any interested individual. The actual technical work of the IETF is done in its working groups, which are organized by topic into several areas. IETF was founded in 1986 and has started working in the Grid domain more recently, mainly involved in the definition of security certificates.

World Wide Web Consortium (W3C) [W3C]

The World Wide Web Consortium is an international consortium that works for the development of Web standards. W3C's mission is to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web. The World Wide Web Consortium (W3C) has been highly successful at promoting a number of proposed standards in various stages of development that are commonly referred to as the "Web services" specifications. The bases for these standards are the HTTP protocol, the XML encoding format, the SOAP remote procedure call mechanism, and the WSDL language.

Web Services Interoperability (WS-I) [WS-I]

WS-I is an open industry organization chartered to promote Web services interoperability across platforms, operating systems and programming languages. The organization's diverse community of Web services leaders helps customers to develop interoperable Web services by providing guidance, recommended practices and supporting resources. All companies interested in promoting Web services interoperability are encouraged to join the effort. Specifically, WS-I creates, promotes and supports generic protocols for the interoperable exchange of messages between Web services. In this context, "generic protocols" are protocols that are independent of any action indicated by a message, other than those actions necessary for its secure, reliable and efficient delivery, and "interoperable" means suitable for multiple operating systems and multiple programming languages.

### 4.3.6 The leading Grid middleware

EGEE, which originally stood for Enabling Grids for E-Science in Europe, now simply stands for Enabling Grids for E-SciencE in recognition of the true international nature of this endeavour. Today this project is the largest collaboration and leading initiative in Grid infrastructure.

EGEE was originally proposed by experts in Grid technologies representing the leading Grid activities in Europe. The project now includes more than 70 project partners organised in twelve partner regions or "federations". Furthermore, with the deployment of the EGEE project structure, several of these partners have begun integrating regional Grid efforts in order to provide coordinated resources to the EGEE project. Today EGEE project has successfully linked 110 sites into the production service. With the new sites, the infrastructure now enables over 10,000 CPUs.

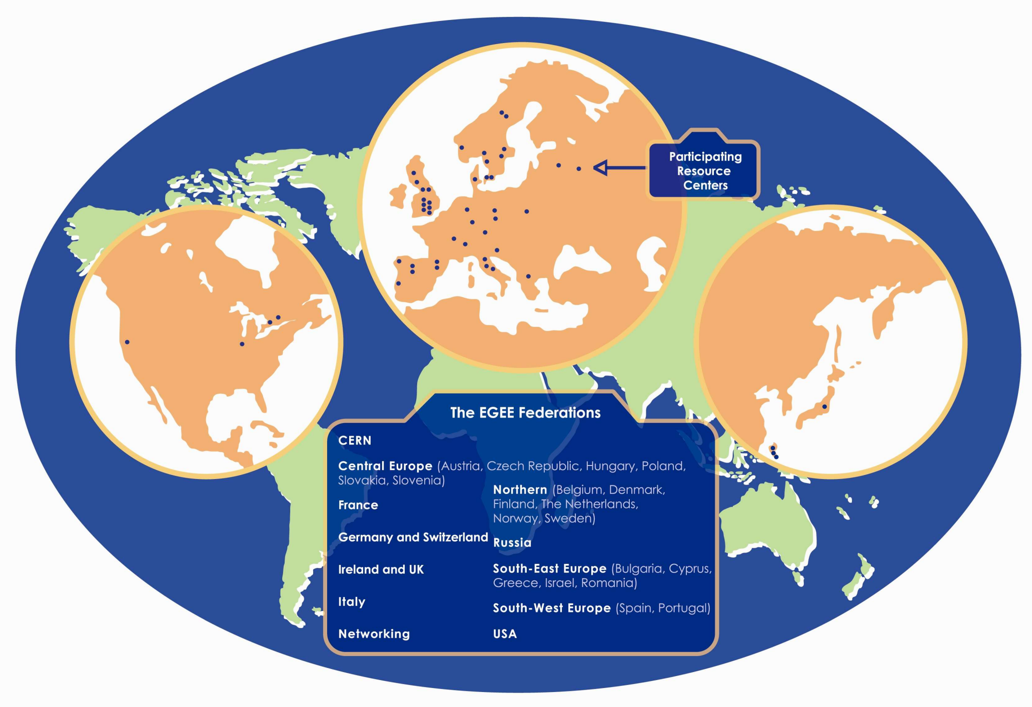



**Figure 2 – The EGEE Federation Map**

The LCG project provides basic resources and infrastructure since 2003. In 2004 Biomedical Grid applications have been deployed on the EGEE production service (i.e., LCG-2). The available resources and user groups are rapidly expanding.

The first phase of the project is expected to deliver by 2006 a production quality Grid infrastructure for the European Research Area and the international scientific community in general, including new challenging applications from 5 different scientific communities, running on the EGEE re-engineered middleware: gLite.

A close collaboration between Europe and USA has been established through EGEE. The purposes of this partnership are:

- The effective exchange of ideas, requirements, solutions and technologies
- The coordinated development of new capabilities
- The establishment of open communication channels
- The joint deployment and testing of middleware
- The early detection of differences and disagreements

The US federation which is officially taking part in the EGEE project includes authoritative representatives of the major Grid initiatives in USA, which are:

- University of Chicago (I. Foster)

- University of Southern California (C. Kesselman)
- University of Wisconsin-Madison (M. Livny)

These institutions are integral part of the gLite technical process, operate gLite prototype in their sites and facilitate exchange of concepts, principals and implementation details between EGEE and OSG (Open Science Grid[4]). As a result, Globus Toolkit, Condor and VDT (Virtual Data Toolkit) developments are oriented to meet EGEE requirements and their adoption in EGEE as part of gLite middleware is facilitated.

[4] The Open Science Grid (OSG) Consortium was formed in 2004 by teams from U.S. universities and national laboratories in order to build and support a production quality peta-scale Grid infrastructure for large scale science. The Open Science Grid will ensure that the U.S. plays a leading role in defining and operating the global Grid infrastructure needed for large-scale collaborative and international scientific research. Reference: http://www.openscienceGrid.org/

# 5. Emerging Scenarios

In this chapter we present three different aspects of the DILIGENT environment will be used when building the DILIGENT testbed and developing the DILIGENT sustainability model (D4.3.2).

In particular we cover Grid and related technologies, including Service Oriented Architectures and Open Source, which form the enabling environment for the DILIGENT application. In fact, other than the potential convergence of Grid services and Web Services even the adoption of a kind of open source for the development and licensing models should be analysed carefully to understand pros and cons that could improve DILIGENT solution.

## 5.1 TRENDS IN GRID TECHNOLOGY

### 5.1.1 Overview

"The original idea of the Grid was ubiquitous access to ones own working environment on any system, from anywhere in the world. The next idea was to make use of the resource potential already available in the internet (all those unused cycles on individual PC's). This idea, although still alive for certain classes of possible applications, does not form the centre of most present Grid-research activities. Most Grid practices focus on the coupling of mid-size systems into a single working environment. Depending on the specific purpose of a Grid research project, the emphasis is on sharing the resources, single sign on, cross system scheduling, extending the Grid software as such, transparency of data or data collections, etc.

In the imagined future, electronic equipment would in due course all be 'Grid aware' once coupled to a network*, i.e.,* it would contain Grid enabling software from the start. Groups of researchers, individuals, or private enterprises would *ad hoc* form a Virtual Organisation to get work done, making use of selected resources from where these are available. Depending on a specific situation accounting and/or budgeting would be in place in some form."[DNA5.1.2]

Grids are still an extremely young product. There are still different concepts of Grids defined at different levels (Data Grids, Computing Grids, Information Grids, Knowledge Grids, etc.) and from different perspectives. Few standards are defined, few stable solutions are available, and lot of work is still being done in the development of these systems.

Nowadays Grids are associated, in some aspects, with several distinct systems:

> ***Like the Web****, Grid computing keeps complexity hidden: multiple users enjoy a single, unified experience.*
>
> ***Unlike the Web****, which mainly enables communication, Grid computing enables full collaboration toward common business goals.*
>
> ***Like peer-to-peer****, Grid computing allows users to share files.*

***Unlike peer-to-peer****, Grid computing allows many-to-many sharing — not only files but other resources as well.*

***Like clusters and distributed computing****, Grids bring computing resources together.*

***Unlike clusters and distributed computing****, which need physical proximity and operating homogeneity, Grids can be geographically distributed and heterogeneous.*

***Like virtualization technologies****, Grid computing enables the virtualization of IT resources.*

***Unlike virtualization technologies****, which virtualize a single system, Grid computing enables the virtualization of vast and disparate IT resources.*

A more rigorous definition of a Grid comes from Gartner Research (March 2003).

"*A Grid is a collection of resources owned by multiple organizations that is coordinated to allow them to solve a common problem. There are three commonly recognized forms of Grid, distinguished by the nature of the resources coordinated:*

*• A computing Grid harnesses multiple computers from several owners to run one very large application problem.*

*• A data Grid uses multiple storage systems from several owners, dividing data across the combined resources to host one very large set of data.*

*• A collaboration Grid ties together multiple collaboration systems from several owners to allow collaboration on a common issue*".

The three types of Grid (computing, data and collaboration) are the "commonly recognized forms of Grid". However, existing Grid-labelled products do not meet the criteria required to be labelled as Grid-products according to the above definitions. Grid computing is a cutting-edge topic in research environments, which business-oriented companies still consider it an immature technology.

Analysts have identified some limitations to the commercial adoption of Grid and consequently the realisation of well-focused Grid-related products.

The greatest inhibitor is organizational borders**.** The cultural challenges of sharing computing across business units and organizations represent the greatest impediment to commercial Grid adoption. Customers note that Grids require a different way of thinking about how to deliver IT datacentre services, and normal resistance to changing behaviour is always the toughest hurdle to overcome in technology adoption.

The general lack of tools and industry standards lead customers to think of Grids as requiring large people and services costs, which would be larger than savings. Vendor investment in management, provisioning, optimization, and scheduling software will help address this market inhibitor.

Content security and confidentiality is also identified as a major inhibitor to broad commercial adoption of Grid technology. Security concerns are also deeply rooted in the cultural and organizational concerns cited above. Security will have to be proven over time to potential customers at a number of levels for Grids to be considered for adoption in shared workload environments.

Accounting is another cultural challenge for both customers and commercial providers. Customers must trust in the provider's accounting system and providers must be able to properly document usage and ensure that proper optimizations have been made. Grid service providers must re-think licensing and find adequate metrics to charge the effective costs for storage and computing services.

The cost of computing resources continues to decline. As the price point for computing resources falls, and the IT industry invests in better tools to remotely manage, deploy, and redeploy hardware and software in a datacentre, many customers will care less about server utilization and over-provisioning as long as the costs of managing that capacity decline. This trend may detract from the promise that Grids bring better utilization server infrastructure and more access to compute resources.

Despite such challenges, attempts to create commercial Grid products are made by leading companies, including Oracle, Sun, IBM, and United devices. In addition, consortia and alliances focusing on specific verticals are formed to promote Grid computing (EGA and Globus consortium. Almost all commercial Grid-initiatives happen in the US, but some European and Japanese initiatives are creating increasing general interests in those geographies.

Current efforts concentrate on creating the necessary impetus for the growth of the required architectures and standards. In order to achieve a worldwide accepted system, lots of work has still to be done and major milestones have to be reached. First, technical communities must finish their work on Grid specifications. Then, these specifications must be ratified and adopted broadly by Grid vendors. Other Grid industry groups must coordinate their work to provide comprehensive Grid reference models, reference implementations, best practices guidelines and interoperability events. All this could take some years, considering the complexity of these new standards (SOA, WSRF, etc).

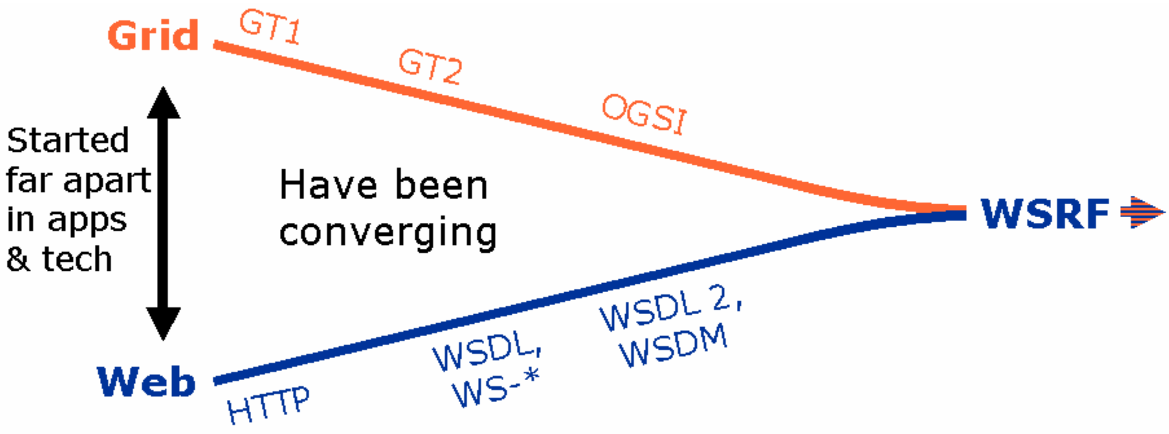
Furthermore, platform vendors must integrate Grid features and standards natively into their products. Finally, Grid won't truly mature until it breaks out of its traditional niche: serving the massive parallel processing needs of supercomputing applications in scientific and engineering environments. Grid potentially could be used to scale and accelerate diverse types of application.

For the rest of this decade, Grid computing will deepen its presence in its traditional scientific and engineering niche. However, Grids will also increasingly penetrate a broader range of commercial environments, thanks to new standards and vendors' growing commitment to this powerful paradigm. [Article1]

### 5.1.2 Standardisation Perspectives

The common standardization effort conducted by the groups described in section 3.3.3. is leading to a convergence between Grid and Web standards. Although these two areas started apart in terms of applications and technologies and in different time periods, they are now approaching. The foreseen result of this convergence is the Web Services

Resource Framework (WSRF). WSRF will allow Grid and Web communities to move forward on a common base (the standard will be defined below).




Open Grid Services Infrastructure (OGSI) Working Group [OGSI]

The WSRF is inspired by the work of the Global Grid Forum's Open Grid Services Infrastructure (OGSI) Working Group. OGSI, which was one of the first attempts to define Web services standards for Grid computing, is a formal specification of the concepts described by the OGSA. It specifies a set of service primitives that define a nucleus of behaviour common to all Grid services. The open source Globus Toolkit® 3.0 (released in mid-2003) contained the first OGSI implementation. Although version 1.0 was released, OGSI was abandoned due to several reasons:

- did not fit well enough with existing Web services architectural principles
- too much stuff in one specification
- too “object oriented”

Based on these problems, WSRF was thought following these guidelines:

- reduce the usage of XML Schema
- partition of OGSI functionality into a family of composable specifications
- explicit distinction between the “service” and the stateful “resources” acted upon by that service

Globus Project is now proposing one WSRF implementation called Globus Toolkit 4.0 (GT4) (other implementations exist, such as WSRF.net or WSRF:Lite). GT4 is a realization of the OGSA requirements and a sort of *de facto* standard for the Grid community while GGF works on standardizing all the different services. Most of these services are implemented on top of WSRF (the toolkit also includes some services that are not implemented on top of WSRF and are called the non-WS components). Globus Toolkit

4, in fact, includes a complete implementation of the WSRF specification. This part of the toolkit (the WSRF implementation) is important since nearly everything else is built on top of it. [IBMStandards], [WSRFStandards], [WSRFFoster]

The most important standards defined by these groups are described below. Other standards can be found in the appendix:

*Simple Object Access Protocol (SOAP) [SOAP] (defined by W3C)*
SOAP is a lightweight XML-based protocol for exchange of information in a decentralized, distributed environment. SOAP can potentially be used in combination with a variety of protocols. However, the only bindings defined in this document describe how to use SOAP in combination with HTTP and HTTP Extension Framework.

*Web Services Description Language (WSDL) [WSDL] (defined by W3C)*

WSDL is an XML format for describing network services as a set of end points operating on messages containing document or procedure-oriented information. WSDL is extensible to allow description of endpoints and their messages regardless of what message formats or network protocols are used to communicate. However, the bindings in this document describe how to use WSDL in conjunction with SOAP 1.1, HTTP GET/POST, and MIME.

*Open Grid Services Architecture (OGSA) [OGSA] (being defined by GGF)*

OGSA defines what Grid services are, and the overall structure and services to be provided in Grid environments. Building on existing Web services standards OGSA defines a Grid service as a Web service that conforms to a particular set of conventions. For example, Grid services are defined in terms of standard WSDL with minor extensions. This is important because it gives us a common and open-standards-based set of techniques to access various Grid services using existing standards, such as SOAP, XML, and WS-Security. With this base, we can add and integrate additional services (such as life cycle management) in a seamless manner. It provides a standard method to find, identify, and utilize new Grid services as they become available. And as an added benefit, OGSA will provide for interoperability between Grids that might have been built using different underlying tools.

*Web Services Resource Framework (WSRF) [WSRF] (being defined by OASIS, GGF, others)*

WSRF is a set of proposed Web services specifications that define a rendering of the WS-Resource approach in terms of specific message exchanges and related XML definitions. These specifications allow the programmer to declare and implement the association between a Web service and one or more stateful resources. They describe the means by which a view of the state of the resource is defined and associated with a Web services description, forming the overall type definition of a WS-Resource. They also describe how the state of a WS-Resource is made accessible through a Web service interface, and define related mechanisms concerned with WS-Resource grouping and addressing.

WSRF is a joint effort by the Grid and Web Services communities. It provides the stateful services that OGSA needs. While OGSA is the architecture, WSRF is the infrastructure on which that architecture is built on. A WSRF goal is to evolve the Grid architecture in a way that's more clearly aligned with the general evolution of Web services. Instead of defining a new type of Grid service, these specifications will allow the services specified in the OGSA to be based completely on standard Web services. [WSRFGT4]

WSRF is composed by several independent specification documents that provide the normative definition of the all framework, between them:

In summary, establishing the Grid and the standards that define it is still very much a "work in progress." The definition and the widespread adoption of standard protocols and interfaces is arguably the most critical problem facing the Grid community today, and for this a large effort is being spent.

Besides these standardization efforts, other guiding efforts are also being made. One example is the e-Infrastructure Reflection Group (eIRG) which is working in the establishment of high level guidelines for Grid computing. The main objective of the eIRG is to support of the political, advisory and monitoring level, the creation of a policy and administrative framework for the easy and cost-effective shared use of electronic resources in Europe (focusing on Grid-computing, data storage, and networking resources) across technological, administrative and national domains [eIRGWebSite].

eIRG members are official delegates from the ministries of Education of the various European countries plus some EC officials. Every six months they produce a document with their recommendation in several areas but always with the main goal of building a healthy European eInfrastructure. These documents can be found on their web site [eIRGWebSitedocs].

### 5.1.3 Substitute solutions

Grid computing has evolved from distributed computing systems. Distributed systems have also evolved (ref. Section 5.2.1) and in some cases offer substitute solutions to Grid technologies. In addition to distributed computing, other solutions that differs even more from Grid computing also exists.

#### Distributed computing

Distributed computing is the process of aggregating the power of several computing entities to collaboratively run a single computational task in a transparent and coherent way, so that they appear as a single, centralized system. Although these systems tend to be specialized systems intended for a single purpose or user group, there are currently several stable solutions which can therefore be seen as a substitute to Grid computing system if we keep in mind their limitations.

#### Cluster Computing

A computer cluster is a group of computers connected by a high-speed network that work together as if they were one machine with multiple CPUs.

Clusters can have different sizes. Scalability is a major advantage of clusters. A cluster can grow simply by adding new PCs to it. Of course there are limits, because somehow the computers have to communicate with each other, which becomes increasingly challenging as the number of computers in a cluster increase.

### Peer to Peer computing

The term “peer-to-peer” (P2P) refers to a class of systems and applications that employ distributed resources to perform a critical function in a decentralized manner.

The resources encompass computing power, data (storage and content), network bandwidth, and presence (computers, human, and other resources). The critical function can be distributed computing, data/content sharing, communication and collaboration, or platform services. Decentralization may apply to algorithms, data, and meta-data, or to all of them. This does not preclude retaining centralization in some parts of the systems and applications if it meets their requirements. Typical P2P systems reside on the edge of the Internet or in ad-hoc networks.

With the pervasive deployment of computers, P2P is increasingly receiving attention in research, product development, and investment circles. This interest ranges from enthusiasm, to disbelief in its potential. Some of the benefits of a P2P approach include: improving scalability by avoiding dependency on centralized points; eliminating the need for costly infrastructure by enabling direct communication among clients; and enabling resource aggregation.

### Internet computing

We will use the SETI@Home application to explain this kind of computing systems. SETI@Home is a virtual "supercomputer" which analyses the data of the Arecibo radio telescope in Puerto Rico, searching for signs of extraterrestrial intelligence. Using the Internet, SETI brings together the processing power of more than 3 million personal computers from around the world, and has already used the equivalent of more than 600.000 years of PC processing power. SETI@Home is a screen-saver program - i.e., it works without impacting normal use of the computer - and any owner of a PC can download it from the Web. The different PCs (the nodes of such Grid) work simultaneously on different parts of the problem, retrieving chunks of data from the Internet and then passing the results to the central system for post-processing. The success of SETI has inspired many other @home applications. [GridCafe]

### Local Grid computing

One solution to solve problems that can be divided into many smaller problems, all independent of each other, is to link computer resources from across a business, a company or an academic institution. The network of computer is then used as a single, unified resource. This solution belongs to the general class of computing called "distributed computing". Many people call this Grid computing, although it fails to meet definitions of Grid Systems that require a Grid to be composed by resources from different organizations.

Local Grid computing makes the most of existing computer resources within an organization. Dedicated software efficiently matches the processing power required by any application with the overall availability. One popular type of software for linking computers in institutions like universities is Condor. Condor is a type of software often referred to as "middleware", because it is not the operating system - the program that runs the computer - nor is it an application program running on the computer, but it is "between" these two, making sure that the application can run optimally on several computers, by automatically checking which computers are available.

Like clusters, local Grid computing is scalable - you can keep adding more PCs and workstations, within reasonable limits. Often the connection between the computers in such a system is a local area network, although it may also be the Internet. Usually the computers are geographically close together, for instance in the same building, and belong to the same administrative domain.

Local Grid computing is limited to a well-defined group of users, a department or several departments inside a corporate firewall, or a few trusted partners across the firewall. Also, such systems typically pool the resources of some dedicated PCs as well as others whose primary purpose is not distributed computing - in other words it involves some "cycle scavenging", at least on a local scale. [GridCafe]

### 5.1.4 Evidences in Grid computing trends

Although any of the presented substitute solutions could be adopted as the appropriate answer for a number of specific needs, Grid computing offers capabilities that extend these other solutions. Therefore Grid could in principle have access to parallel computers, clusters, farms, local Grids, even Internet computing solutions, and would choose the appropriate tool for a given calculation. In this sense, the Grid is the most generalised, globalised form of distributed computing.

Grid computing can be seen as the evolution of local Grid computing to the global scale, made possible by the advent of very high-speed Internet connections and powerful computer processors that are able to run quite complex middleware as background tasks without interrupting other tasks the computer is trying to handle. As Internet quality and bandwidth increases, the differences between having two PCs in the same office, the same building, the same city or the same country grows smaller.

By developing sophisticated middleware which makes sure widely distributed resources are used effectively, Grid computing gives the user the impression of shrinking the distances further still. Furthermore, as the middleware gets more sophisticated, it can deal with the inevitable differences between the types of computers that are being used in a highly distributed system that span multiple organizations.

Grid computing focus is on large scale sharing, which goes beyond institutional boundaries. Also, Grid computing leans more to using dedicated systems, such as scientific computer centers, rather than cycle scavenging. Finally, and what is in some ways the most challenging aspect, Grid computing aims to use resources that are not centrally controlled. The sharing is across boundaries - institutional and even national - which adds considerable complexity, while bringing also huge potential benefits. Grid computing must satisfy three criteria:

- no central administrative control of the computers involved (eliminates clusters, farms, and local Grid computing)
- Use of general-purpose protocols (eliminates single-purpose systems such as SETI@Home)
- High quality of service (eliminates peer-to-peer and means that Grids should not rely on cycle scavenging from individual processors, but rather on load balancing between different independent large resources, such as clusters and local Grids)

The Grid is clearly a "hot topic" in the ITC environment. It offers a solution to the problem of massive demand for processing power and data storage of modern 21st century scientific applications. The well known World Wide Web came from the same mould and from similar needs for sharing of information. The Grid, like the Web, comes from requirements of cutting edge scientific experiments and research fields. It is hard to see or imagine how exactly the general public will benefit directly from a multi-domain production Grid. However, as for the Web, the availability of a trusted computing Grid will undoubtedly stimulate the imagination of young, talented and visionary computer scientists, in developing a new generation of Grid-enabled applications. The timescale and impact is open for debate, while the scientific community leaps ahead, using the current Grid to explore new corners of science for the benefit of humanity.

## 5.2 RELATED TECHNOLOGY TRENDS

Findings from analysing EU and non-EU projects, as described in the project analysis section of the document, are that the following emerging Information and Communication Technologies (ICT) paradigms and approaches are dominating the evolution of the technological scenario in e-Science, both in terms of application performances and in terms of services delivered:

1. Service-oriented architectures (SOA):
   - IT infrastructures involving multi-players collaboration, employing a service-oriented approach;
   - enabled by using Web Services;
   - focus on services to be delivered to the user, rather than a content-oriented approach.
2. Open infrastructures, which allow significant cost reduction for developing and deploying new services. Anyone can contribute to software development, thus maximising the re-use of existing results; open infrastructures are implemented from the open source model and require that software is distributed through Open Source licences.

As previously mentioned, among its different goals, DILIGENT also aims at advancing the state of the art in the Digital Library sector by:

1. employing a service-oriented approach, focusing on the right added value for the user;
2. relying on a Grid infrastructure, in order to reduce the cost of development and deployment of new services by maximising the re-use of existing results, and to reduce maintenance costs by eliminating the need for a centralised organisational and/or technical infrastructure.

In order to better achieve the above mentioned goals, DILIGENT will deal with a keen exploitation of such emerging ICT paradigms, as they are dominating the evolution of the worldwide technological scenario, both in terms of application performances and in terms of services delivered.

### 5.2.1 Service Oriented Architecture (SOA)

Over the last years, the field of software architecture has attempted to deal with increasing levels of software complexity. Keywords are “heterogeneity” and “application integration”. Systems must accommodate an endless variety of hardware, operating systems, middleware, languages and data stores, which changes focus from the content to service.

Each shift, in the evolution of programming models, was made to the deal with greater levels of software complexity and to enable the assembly of applications through parts, components, or services.

In this scenario, service-oriented architectures allows for designing software systems that provide services to other applications through published and discoverable interfaces, and where services can be invoked over a network.

When implementing a service-oriented architecture, e.g., using Web services technologies, a new way of building applications will be created within a more powerful and flexible programming model. Development and ownership costs as well as implementation risks are reduced. SOA is both architecture and a programming model, a way of thinking about building software.

### 5.2.2 Open Infrastructure

Open Infrastructures are system implementing Open Source concepts of free access to information and software code for social improvement.

The basic idea behind open source is very simple: when programmers can read, redistribute, and modify the source code for a piece of software, the software evolves. People improve it, adapt it and fix bugs.

In the open source community this rapid evolutionary process can produce better software than the traditional closed model, in which only a very few programmers can see the source and everybody else must blindly use an opaque block of bits.

Stallman defines software with four freedoms:

- freedom to use
- freedom to read the source code
- freedom to modify and improve the source
- freedom to redistribute originals and modifications.

Open Source Initiative (OSI) maintains a list of open source licenses that conform to the Open Source Definition, have been through public scrutiny, and have been approved by them. The "classic" licenses are GPL, LGPL, BSD, MIT and Mozilla. Many other licenses have been submitted for review and approval by OSI.

Here below are mentioned the main benefits produced by the open source to distinct categories of user/stakeholders:

Business users

- help many companies and individuals to collaborate on a software that none of them could achieve alone, sharing costs and efforts;
- it is the rapid bug-fixes and the changes that the user asks for, done to the user's own schedule;
- it helps to increase security, because software code is in the public view and it will be exposed to extreme examination, with problems being found and fixed instead of being kept secret until the wrong person discovers them;
- it represents a chance to fight against monopolistic market politics;
- it allows to increased reliability

Customers

- customers stop to be prisoners of Big software vendors, being totally at the mercy of their pricing policies, their willingness to fixed bugs, their strategic decisions
- customer can survive the collapse of software vendors, controlling their own IT projects and investments

Although Open Source is growing in acceptance, unclear business models and the risk introduced by lack of accountability from vendors represent major inhibitors to the Open Source movement.

## 5.3 MARKET TRENDS

### 5.3.1 Demand for environment enabling collaboration and Knowledge exchange

In parallel with the evolution of Digital Libraries systems we are observing a large expansion in the demand for Digital Libraries. Research work today, in any field, is often a collaborative effort carried out by groups belonging to different organizations worldwide. Motivated by a common goal and funding opportunities, these groups dynamically form virtual research organizations that share resources e.g., knowledge, experimentation results, instruments, for the duration of their collaboration, creating new and more powerful virtual research environments. These virtual research organizations, set-up by individuals that do not necessarily have a great economic power and technical expertise more and more frequently require Digital Libraries as tools for accelerating their research activities. This potential user group demands less expensive and dynamic Digital Libraries development models. They want to be able to set up new Digital Libraries that serve their needs for the duration of their collaborations in an acceptable time frame and with an acceptable cost.

The current Digital Libraries development model cannot satisfy this large demand and a radical change is needed if we want to be able to address the new and emerging requirements. New technologies must be investigated to support the implementation of the novel functionalities on the more versatile digital information objects. New organizational and development models must be introduced to reduce their cost of and to speed up their realization time.

### 5.3.2 Demand for Grid Technology

In April 2004 the Economist Intelligence Unit conducted a survey [EconomistSurvey] on Grid computing with 177 senior executives. The selected executives represent several areas of industry and services, such as financial services, telecoms, education, transport, government and others. The following are extracted from the results of the survey:

**Has your company implemented Grid computing?**
- Yes 9
- No 91

**If you have implemented Grid computing, how satisfied are you with the outcome?**
- Extremely satisfied 11
- Satisfied 84
- Dissatisfied 5
- Extremely dissatisfied 0

**How would you characterize your knowledge of Grid computing?**
- Excellent 0
- Good 12
- Average 27
- Fair 19
- Poor 42

**Which of the following options best describes your company's attitude to Grid computing?**
- We were early adopters 6
- We are now deploying/about to deploy this technology 8
- We are evaluating the case for Grid computing 37
- We are unlikely to implement Grid computing in the foreseeable future 44
- We have decided not to implement Grid computing 6

**How big an impact do you believe Grid computing will have on your industry in the next five years?**
- Massive impact 4
- Significant impact 33
- Average impact 30
- Small impact 26
- No impact 7

The outcome of the survey suggests that Grid computing is still not massively implemented in the market and that knowledge of Grid systems is, in most cases (more than 50%), lower than average. On the other hand, it seems that a significant number of companies are evaluating Grid systems (37%). It is also important to highlight that 66% of the executives believe Grid computing will have an average or significant impact on their industry. Therefore we can predict that the demand of Grid systems will increase in the next years and companies will try to get more involved.

Grid computing systems can be the answer to various activities and application needs, therefore someone who pretends to compute or share some sort of information can be a possible Grid buyer. The interest on Grid technology is increasing and new communities of possible buyers are “discovering” how Grid can help them in their activities. Areas in which interest in Grid computing has already been raised are [GridStart, Article2]:

Research Collaboration

Research-oriented organizations and universities practicing in advanced research collaboration areas require the analysis of tremendous amounts of data. Some examples of such projects are subatomic particle and high energy physics experiments, remote sensing sources for earth simulation and modelling, and analysis of the human genome sequence. These virtual organizations engaged in research collaboration activities generate petabytes of data and require tremendous amounts of storage space and thousands of computing processors. Researchers in these fields must share data, computational processors, and hardware instrumentation such as telescopes and advanced testing equipment. Most of these resources are pertaining to data-intensive processing, and are widely dispersed over a large geographical area. The Grid computing discipline provides mechanisms for resource sharing by forming one or more virtual organizations providing specific sharing capabilities. Such virtual organizations are constituted to resolve specific research problems with a wide range of participants from different regions of the world. This formation of dynamic virtual organizations provides capabilities to dynamically add and delete virtual organization participants, manage the "on-demand" sharing of resources, plus provisioning of a common and integrated secure framework for data interchange and access.

#### Life sciences

Current major efforts surrounding genomic, proteomics and molecular biology provides the basis for many of the current Grid computing advancements. The Grid computing systems can provide a common infrastructure for data access, and at the same time, provide secure data access mechanisms while processing the data. Therefore life sciences can utilize Grid computing systems to execute sequence comparison algorithms and enable molecular modelling using the above-collected secured data. This now provides the Life Sciences sector the ability to afford world-class information analysis respective to this discussion, while at the same time providing faster response times and far more accurate results.

#### Medicine

Medicine is one of the candidates for the early adoption of the Grid. Several fields such as computer based drug design, medical imaging, medical simulations, analysis of 2D, 3D, 4D images, simulation (radiotherapy, plastic surgery) and cancer diagnosis are growing extremely fast. As a result of their work, large amount of data in heterogeneous sources and formats are becoming available. These facts are increasing the demand for large computational power and easy accessibility to heterogeneous data sources in these health domains.

#### Financial Markets

Financial Markets are complex and very quickly changing environments in which companies carry out their business activities. To take advantage of opportunities in these markets it requires precise evaluation of risks and of the suitability of the products with respect to customers' demands. Financial companies evaluate potential products using complex numerically intensive algorithms and applications which analyze large amounts of data. Requirements related to sophistication, accuracy and faster execution are among the more salient objectives across financial communities. These objectives can be achieved by real-time access to the current and historical market data, complex financial modelling based on the respective data, and faster response times to user queries. Grid computing provides the financial analysis and services industry sector with advanced systems delivering all the competitive solutions in Grid computing. These solutions exemplify the infrastructure and business agility necessary to meet and exceed the uniqueness that the financial analysis and services industry sector requires. This particular value statement is accomplished by the fact that many of these solutions in this industry are dependent upon providing increased access to massive amounts of data, real-time modelling, and faster execution by using the Grid job scheduling and data access features. For this to be most successful, these financial institutions tend to form virtual organizations with participation from several different departments and other external organizations. In addition to the use of existing resources, a Grid system can provide more efficiency by easily adapting to the rapidly changing algorithms pertaining to the financial analytics.

#### Engineering and Design

The enormous competitive pressure in the business and industry sectors today affords most engineering and design far less turnaround time. They need mechanisms to capture data, speed up the analysis on the data, and provide faster responses to market needs. As we already know, these engineering activities and design solutions are inherently complex across several dimensions, and the processing requirements are much more intense than that of traditional solutions of the past. Grid computing systems provide a wide range of capabilities that address the above kinds of analysis and modeling activities. These advanced types of solutions also provide complex job schedulers and resource managers to deal with computing power requirements. This enables automobile manufacturers (as an example) to shorten analysis and design times, all while minimizing both capital expenditures and operational expenditures.

#### Collaborative Games

There are collaborative types of Grid computing disciplines that are involving emerging technologies to support online games, while utilizing on-demand provisioning of computation-intensive resources, such as computers and storage networks. These resources are selected based on the requirements, often involving aspects such as volume of traffic and number of players, rather than centralized servers and other fixed resources. These on-demand-driven games provide a flexible approach with a reduced up-front cost on hardware and software resources. We can imagine that these games use an increasing number of computing resources with an increase in the number of concurrent players and a decrease in resource usage with a fewer number of players. Grid computing gaming environments are capable of supporting such virtualized environments for enabling collaborative gaming.

#### Government

The Grid computing environments also focus on providing coordinated access to massive amounts of data held across various agencies in a government. This provides faster access to solve critical problems, such as emergency situations, and other normal activities including intelligence analysis. These key environments provide more efficient decision making with less turnaround time. Grid computing enables the creation of virtual organizations, including many participants from various governmental agencies (e.g., state and federal, local or country, etc.). This is necessary in order to provide the data needed for government functions, in a real-time manner, while performing the analysis on the data to detect the solution aspects of the specific problems being addressed. The formation of virtual organizations, and the respective elements of security, is most challenging due to the high levels of security in government and the very complex requirements.

# 6. Market Composition

DILIGENT aims at developing an advanced test-bed by integrating Grid and digital library technology for different Virtual Research Organisations enabling the offering of technology-based services with the main purpose of:

- access shared knowledge, and
- collaborate to reach common research objectives.

The composition of the DILIGENT market emerges as very rich and heterogeneous in terms of nature and typology of actors.

## 6.1 MARKET DESCRIPTION

The LIBECON project - funded by the DG for the Information Society of the European Commission within the 5th FP for Research –developed a continuously updated database of statistics about library activities and associated costs in the countries of the "enlarged EU"[5] and in Japan and the USA. An internet site (www.libecon.org) has been established to collect and make available library statistics with the purpose of describing progress and trends in library services.

Libraries are developing an ever increasing role in the supply of knowledge, making rapid progress in methods of distributing knowledge by electronic means.

Here below are indicated some relevant macro numbers for all library sectors (national, public, school, tertiary education and special) in the **Enlarged EU**, as of 2001.

- Registered users were almost 138 million, 30% of the population between 1997 and 2001

The number of register is very high and in the enlarged EU the total has increased by more than 2% since 1997. Although more and more citizens are registering as users, the trend is that the number of loans and visits is decreasing. What seems to be happening is that electronic services are attracting users to use libraries, but not always through physical visits and check outs of books.

- Spending on libraries was more than 14 million €

Expenses, that include staff expenditure, stock or collection acquisition, electronic information and automation (software and hardware) expenditure, have increased by more than 22% since 1997.

- Full time equivalent staff were 336,673

[5] By the term enlarged EU we mean the 15 existing EU countries (Austria, Belgium, Denmark, Finland France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, UK) plus the 10 acceding ones (Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, the Slovak Republic, and Slovenia).

Total staff numbers have increased slightly, though in the EU they have decreased slightly on a per thousand population basis. The Accession countries have much more staff than the EU on a per capita basis and there is no evidence of the gap closing.

- Number of service points were 186,826

Of the above totals, about 70,000 in the enlarged EU are public library service points, about 37% of all service points (service points include main library, branch and mobile library). The total number of service points is gradually decreasing and the rate of decrease is greater in the Accession countries. In these countries the incidence of service points per 1000 population is almost 3 times that in the EU. It is expected that this trend will continue. Average spending per service point in the Enlarged EU was about 73,000 Euros in 2001 and average staff per service point was about 1.8 full time equivalents.

- Number of visits 3,171,215,882, with 7.0 visits per head

Visits in the EU are gradually declining whereas in the Accession countries they are increasing from a low base.

Availability of electronic services impact both ways on this measure: they draw people to the library to use e.g., internet enabled PCs, and addition to allowing them to obtain information remotely. Information on the use people make of the library when they visit is not generally available on an internationally comparable basis.

- Number of loans 3,324,238,175 with 7.3 loans per head

Loans are decreasing slightly both in absolute numbers and per head of population. So far this traditional metric of library use has not been enhanced with a metric for use of electronic services or use at a distance, which could be aggregated nationally and internationally. Though from other sources we know that the availability and use of such services is widespread and increasing, it remains statistically invisible.

- Number of books and bound periodicals were more than 2.5 billion

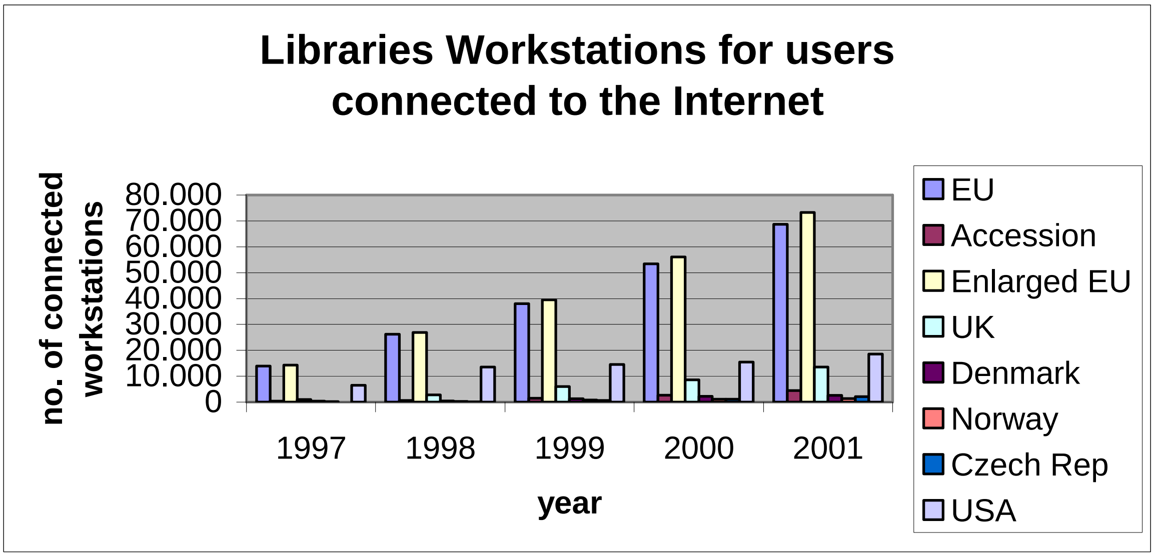
Collections in the major traditional media continue to be maintained. Over this period there has also been a trend towards digitisation of older materials and towards legal deposit of “born digital” materials which has not been captured statistically.

- Number of libraries’ workstations connected to the internet 73.206 and increasing rapidly

This question was answered by 10 European countries in 2001 and the grossed up totals are shown in the tables below, but data from the USA are also available.

| | 1997 | 1998 | 1999 | 2000 | 2001 | Increase 1997-2001 |
|---|---|---|---|---|---|---|
| EU | 13.908 | 26.259 | 38.004 | 53.379 | 68.730 | 494% |
| Accession | 336 | 618 | 1.492 | 2.729 | 4.476 | 1332% |
| **Enlarged EU** | **14.244** | **26.877** | **39.496** | **56.108** | **73.206** | **514%** |
| UK | 1.000 | 2.800 | 6.000 | 8.600 | 13.500 | 1350% |
| Denmark | 344 | 463 | 1.271 | 2.218 | 2.551 | 742% |
| Norway | 154 | 292 | 841 | 1.072 | 1.369 | 889% |
| Czech Rep | 45 | 81 | 609 | 1.083 | 2.116 | 4702% |
| USA | 6.483 | 13.526 | 14.524 | 15.521 | 18.534 | 286% |

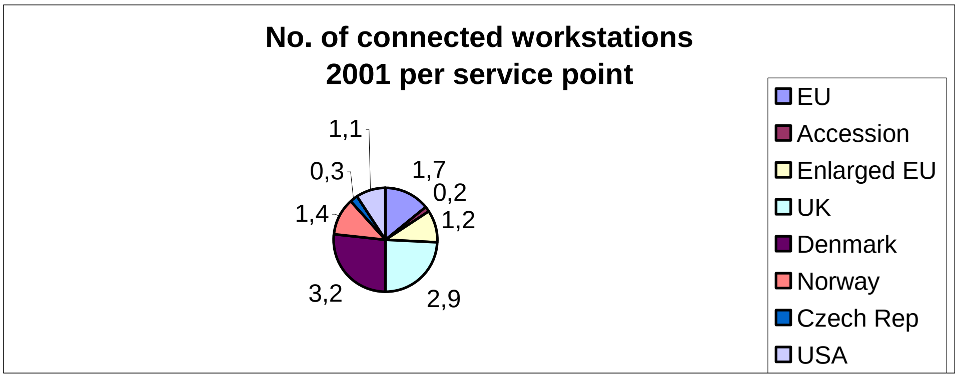
Libraries' workstations connected to the internet from 1997 to 2001



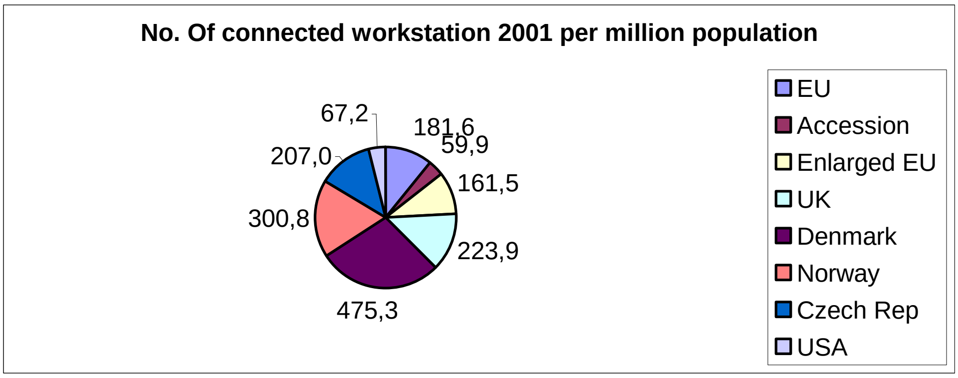

In the following table we show some ratios, which demonstrate the results of the trends as they appeared to be in 2001:

| | No. of connected workstation 2001 | No. of connected workstation 2001 per service poit | No. of connected workstation 2001 per million population |
|---|---|---|---|
| EU | 68.730 | 1,7 | 181,6 |
| Accession | 4.476 | 0,2 | 59,9 |
| **Enlarged EU** | **73.206** | **1,2** | **161,5** |
| UK | 13.500 | 2,9 | 223,9 |
| Denmark | 2.551 | 3,2 | 475,3 |
| Norway | 1.369 | 1,4 | 300,8 |
| Czech Rep | 2.116 | 0,3 | 207,0 |
| USA | 18.534 | 1,1 | 67,2 |

- Whilst the Enlarged EU average of numbers of workstations connected per service point is 1.2, this is greatly exceeded in the UK and Denmark. This measure is affected by the average size of service points (as measured by population served). In Denmark and the Czech Republic, the population served per service point is below the EU average, whereas in the UK it is well above it.




- The measure number of workstations connected (in public libraries) per million population perhaps gives the best indication of the extent of availability of this service. On this measure the Czech Republic comes close to matching the UK and massively outperforms the USA but an Internet connection in only one library in 3 still excludes many people from this service. There are massive disparities and this is an area of rapid development in some countries. It is to be hoped that the disparities will be addressed shortly, assisted by national and regional programmes within the framework of the e-Europe targets for 2005.




Trends on the above mentioned data encourage the development of projects on Digital Libraries. However, with reference to IT based services provided by Digital Libraries, there are not available data. The Benchmarking analysis on Digital Libraries websites has highlighted the most delivered services by advanced examples of Digital Libraries (see benchmarking Grid on appendix).

## 6.2 Digital Libraries Market needs

In the digital era, the main strategic need for DL owners is to guarantee their financial **sustainability** through making **digital asset management** a core function.

Since users **increasingly expect digital services**, DL owners consider digital asset projects as vital to the success of the following:

- **service delivered**, including the quality of service (QoS) and the delivery process;
- **management process**, including both internal processes management and external marketing and selling activities.

The following table summarizes the main internal and Digital libraries external activities executed by e-Science institutions or organisations.

| **Activities and Services** | | | | |
|---|---|---|---|---|
| **Access and Dissemination** | **Education** | **Building Community** | **eTransaction Trustiness** | **Process Management** |
| Access and description of catalogues/collections<br><br>Search & Browse<br><br>Personal collections<br><br>Annotation<br><br>Metadata management | eLearning activities<br><br>Edutainment activities | Chat, Forum<br><br>Help-desk<br><br>Peer-review<br><br>Recommender | DRM services<br><br>Support for ePayment | Sharing documents<br><br>Loans management<br><br>Objects comparison<br><br>Benchmarking<br><br>Tools for acquirement and preservation |

**Summary of activities and services delivered by e-Science institutions**

It is important to observe that the category of activities enabling the creation of new content, such as document self-archiving and self-publishing, collaborative document creation, creation of annotation remain specific activities and services for Digital Libraries and are not extended to virtual museums.

Nowadays, e-Science institutions need to execute such mentioned activities by effectively delivering the above listed services over the web.

Further, the digitization of resources and the execution of internal and external management services open up for new modes of access, delivery and service usage. This also opens for a larger potential audience and gives final users new opportunities to experience e-Science.

The main needs that Libraries attempts to meet with the adoption of ICT with regarding providing digital content services include:

| Needs of Libraries |
|---|
| *Access and Dissemination* |
| To increase **easy and immediate access:**<br>- to high demanded and frequently used items and materials held remotely<br>- to individual components within items (e.g., articles within journals)<br>- to collections of items |
| To **enhance search**, including full text and natural language, based on metadata. |
| To virtually **aggregate** dispersed collections, also into personal collections of multimedia objects. |
| To **enhance** digital images in terms of size, sharpness, colour, noise reduction, etc. |
| To **display** materials which are in **inaccessible** formats, for instance, large volumes or maps. |
| To display the user's **geographical location of resources** when searched |
| *Education* |
| To **increase learning effectiveness** providing with interactive virtual learning environments (VLE)<br>To **integrate** different media (images, sounds, video etc), into teaching materials |
| *Building Community* |
| To develop **programs and activities** for creating **loyalty** among community members<br>To develop **customized** services and communication |
| *e-Transactions and Trustiness* |
| To enable resource selling, dealing with IPR and copyright management |
| To operate in trusted networks with **authenticated** processes and transactions |
| *Process Management* |
| To **reduce** the load or cost of delivery and the **price** of delivery |
| To **increase** the effectiveness of services delivery |
| To **improve** the effectiveness of both internal and external processes management exploiting and adopting best practices developed by other e-Science Institutions or Organizations |

**Summary of needs by e-Science Institutions**

New media and publishers' needs

New Media Agencies (including Architects and Graphic designers), Film makers, Advertisers, Editors, and also Journalists and Writers are mainly interested in digital images and pictures. They usually expect web based services that offer:

- Original support of the image
- Highest quality of both negative and digital image
- Collections of rare images
- Choice between free and pay-per-use downloading of images/pictures
- Copyright cleared transaction through a licence agreement according to their individual usage

Educational sector' needs

Schools, universities and post-graduates courses, but also teachers and students usually require the opportunity to get use of the previously listed services for individuals in tourism, but also:

- e-Learning activities through Virtual Learning Environments (VLE)
- fast access to required resources with very short wait times (also for downloading), consistent with current best practice;
- ability to make e-loan with libraries;
- ability to subscribe to e-journal on specific research areas;
- requirement to be informed on the latest news on their specific field
- ability to find rare images/pictures related to their research field

## 6.3 MARKET SEGMENTATION

The purpose of the market segmentation analysis is to identify differences and similarities between groups of customers and users [Johnsen].

A wide group of different users and organizations has been analyzed with the aim of partition the market for Digital Libraries into meaningful segments.

The following characteristics have been used to segment the market for Digital Libraries:

1. User group
2. Usage situation
3. Size of solution
4. Dynamic or static content and configuration requirements
5. Whether Digital Libraries are business critical
6. Budget
7. Geographic location of customers
8. Service or license delivery

An analysis of different user groups is summarized in Table 6.1.

## Table 6.1 Market segmentation analysis

| User group | Usage | Sol-ution size[6] | Compu-tational intensity | Dyn-amics[7] | Business critical | Budget | Selection criteria | Geo-graphy | Ser-vice[8] | Example users |
|---|---|---|---|---|---|---|---|---|---|---|
| e-Science | Research and dissem-ination | Large/ mediu m | High | Dynamic | Yes | Large | Unknown | EU/ US/ Japan | No | ESA<br><br>CERN |
| eGovern-ment | Citizen self service, research, dissem-ination | Large/ mediu m | Low | Static | Possibly | Large | Unknown | EU/ US/ APAC | Yes | EU<br><br>National govern-ments |
| eHealth | Research | Large/ mediu m | High | Dynamic | Possibly | Large | Unknown | EU/ US/ Japan | Yes | Research institutions |
| eLearning | Retrieval and dissemi-nation | Mediu m/ small | Low | Dynamic | No | Small | Unknown | EU/ US/ Japan | Yes | eLeraning providers |
| ECulture | Preseravation and accessibility | Large/ mediu m | Low – but high Storage and bandwith requireme nts | Dynamic | Yes | Small | Unknown | EU/US | Yes | Bricks community<br><br>Dspace,<br><br>Digital Memory |
| Conf-erences | Dissemi-nation | Small | Low | Dynamic | No | Small | Unknown | EU/ US/ Japan | Yes | Conference organizers |
| Media | Retrieval and dissemi-nation | Large/ mediu m/ small | Low/ High | Dynamic | Yes | Large/ small | Unknown | EU/ US/ Japan | Yes | RAI |
| Libraries | Retrieval and dissemi-nation | Large/ mediu m | Low | Static | Yes | Small | Unknown | EU/ US/ Japan | Yes | Large libraries |

Based on the above analysis, the following observations can be made:

- There are user groups where member have money and to whom Digital Libraries are business critical
- Both public and private entities has a demand for Digital Libraries functionalities
- Most entities have data of dynamic nature
- Most groups have members all over the industrialized world

[6] Based on amounts of data or computational requirements. L= Large, 2= Medium, S= Small
[7] Change frequency of content and configuration.
[8] Indicates whether customer is likely to demand digital library solution to be delivered as service.

- Some user groups have a demand for both small and large scale Digital Libraries solutions

- Some user groups have large data volumes and a need for lots of computational resources

Using the above analysis we find it useful to segment the market into the following relevant groups:

**Table 6.2 Digital libraries market segments**

| Market segment | User group members | Segment characteristics | Comments |
|---|---|---|---|
| Libraries | Public and corporate libraries | - Digital libraries are business critical<br>- Small budgets<br>- Large amounts of data and low computational requirements | Familiarity with Digital Libraries, low IT competency. |
| e-Science | Private and public research organizations | - Digital libraries are business critical<br>- Large budgets<br>- Large amounts of data and high computational requirements<br>- Requires different types and sizes of Digital Libraries | Familiar with Digital Libraries, high IT competence, drives Grid development |
| Media/publishing companies | Private media companies and public broadcasters | - Large amounts of data<br>- Unknown budgets | Demand for Digital Libraries is questionable. |

# 7. DILIGENT Competences

The DILIGENT consortium is composed of scientific, academic and commercial partners whose joint competences span a wide range of research, development and delivery capabilities.

This following table provides a high level overview of DILIGENT's joint competences:

| FUNCTIONAL AREA | CAPABILITY |
|---|---|
| Research and Development | Expertise in DL<br>Expertise in Grid<br>Expertise in search and information retrieval<br>Effective collaboration with EGEE<br>Effective collaboration with user groups<br>Alignment with open-source community |
| Solutions and Services Delivery: | Engagement of industrial partners<br>Understanding of market requirements |
| Dissemination and Marketing | Promoting and exploiting innovative and pragmatic solution<br>Understanding of and responsiveness to market technology trends |
| Finance | Effective financial management (with public financing)<br>Expertise in coordinating multiple partners<br>Expertise in responding to EC requirements |

In summary, DILIGENT's capabilities are the strongest within research and development compared to sales, marketing and product delivery.

# 8. Conclusions

In writing this report we have attempted to identify and analyse factors influencing DILIGENT's sustainability strategy. Our work will form the basis from which such a strategy is to be developed in D4.3.2. This document therefore contains few conclusions, as those are to be drawn in the next phases of the sustainability strategy process.

The sections below summarizes critical success factors that are of relevance to DILIGENT's sustainability strategy, and a SWOT analysis – which attempts to juxtapose internal and external factors influencing DILIGENT going forward.

## 8.1 Critical Success Factors

Following the observations and analysis in the previous sections, we have identified the following factors as most important in order for DILIGENT to succeed in delivering value to users:

1. Continued funding
   The Digital Libraries market is immature as far as user requirements and product maturity is concerned. We therefore expect that achieving the vision of cost efficient and flexible Digital Libraries on Grid infrastructure will require external funding for several years to come in addition to the revenue opportunities that are to be identified in D4.3.2.

2. Delivering the right functionality at the right time
   This is a common success criteria for all product development organizations. Specifically for DILIGENT, it means the ability to identify the most attractive market segments in the Digital Libraries space and focus on delivering features that tightly match user requirements.

3. Quick wins
   DILIGENT must be able to demonstrate with real world examples that it is meaningful to Digital Libraries on Grid infrastructure.

4. Technology pragmatism
   Due to the evolving nature of Grid technology, DILIGENT must be pragmatic in using alternative technologies for a migration path to a Grid-solution only.

## 8.2 SWOT Analysis

A SWOT analysis assists us in making the distinction between DILIGENT's external environment and DILIGENT's internal environment.

Strengths and Weaknesses refer to the "Internal environment" in a relative (to competition), manner not absolute.

Opportunities and Threats refer to the "External environment" and therefore involve an assessment of risks.

| Strengths | Weaknesses |
| --- | --- |
| - EU DILIGENT project funding (i.e., public funding)<br>- Cross discipline competences (i.e., Grid and Digital Libraries)<br>- Close cooperation between technology providers and users | - Research and commercial partners with conflicting objectives and expectations regarding commercialization of project results<br>- Unproven Grid technologies |
| **Opportunities** | **Threats** |
| - Cost efficient enabler of Digital Libraries solutions<br>- Adoption of the DILIGENT developed technology in other application areas | - Grid technology unable to meet customer expectations<br>- Large initial investments required to establish Grid infrastructure<br>- Small market for Digital Libraries outside academia |

Grid technology market requirements are not covered, since the focus of DILIGENT is to improve Digital Libraries creation and management with Grids as an enabling technology. Potential impacts of DILIGENT results in the Grid market will be taken into account as soon as this becomes apparent.

# 9. Appendix: Grid standards

*WS-Addressing*

(being defined by IBM, BEA, Microsoft)

WS-Addressing provides transport-neutral mechanisms to address Web services and messages. Specifically, this specification defines XML elements to identify Web service endpoints and to secure end-to-end endpoint identification in messages. This specification enables messaging systems to support message transmission through networks that include processing nodes such as endpoint managers, firewalls, and gateways in a transport-neutral manner.

*WS-Notification*

(being defined by IBM, Akamai Technologies Inc, Globus Alliance, Hewlett-Packard, SAP AG, Sonic Software Corporation, Tibco Software)

WS-Notification implements the notification pattern, where a service provider, or other entity, initiates messages based on a subscription or registration of interest from a service requestor. It defines how the publish/subscribe (pub sub) pattern commonly used in Message-Oriented middleware products can be realized using Web services. This includes brokered as well as direct pub sub which allows the publisher/subscribers to be decoupled and provides greater scalability.

*WS-PolicyFramework*

(being defined by IBM, BEA, Microsoft, SAP, Sonic Software, VeriSign)

The Web Services Policy Framework defines a general purpose model and corresponding syntax to describe and communicate Web services policies so that Service consumers can discover the information they need to know to be able to access services from a Service Provider.

*WS-ReliableMessaging*

(being defined by IBM, BEA, Microsoft, TIBCO)

This specification describes a protocol that allows messages to be delivered reliably between distributed applications in the presence of software component, system, or network failures. To support interoperable Web services, a SOAP binding is defined within this specification.

*WS-ResourceLifetime*

(being defined by IBM, Globus Alliance, HP)

WS-ResourceLifetime defines two ways of destroying a WS-Resource: immediate and scheduled. This allows designers flexibility to design how their Web services applications can clean up resources no longer needed.

*WS-ResourceProperties*

(being defined by IBM, Globus Alliance, HP)

WS-ResourceProperties defines how the data associated with a stateful resource can be queried and changed using Web services technologies. This allows a standard means by which data associated with a WS-Resource can be accessed by clients.

*WS-SecureConversation*

(being defined by IBM, BEA, Computer Associates, Layer 7 Technologies, Microsoft, Netegrity, Oblix, OpenNetwork Technologies, Ping Identity Corp, Reactivity, RSA Security, VeriSign, Westbridge Technology)
The Web Services Secure Conversation Language is built on top of the WS-Security and WS-Policy models to provide secure communication between services. WS-Security focuses on the message authentication model but not a security context, and thus is subject several forms of security attacks. This specification defines mechanisms for establishing and sharing security contexts, and deriving keys from security contexts, to enable a secure conversation.

*WS-Security*
(being defined by OASIS)
Web Services Security (WSS) offers a trusted means for applying security to Web services by providing the necessary technical foundation for higher-level services. WSS builds upon existing security technologies such as XML Digital Signature, XML Encryption and X.509 Certificates to deliver an industry standard way of securing Web services message exchanges. Providing a framework within which authentication and authorization take place, WSS lets you apply existing security technology and infrastructure in a Web services environment.

*WS-Security Addendum*
(being defined by IBM, Microsoft, VeriSign)
This addendum to the WS-Security specification clarifies elements released in the original document and introduces some new items including timestamps, and passing around passwords and security certificates.

*WS-Security Kerberos Binding*
(being defined by  IBM, Microsoft)
This document describes how to use Web services Security Specifications with Kerberos.

*WS-SecurityPolicy*
(being dfined by IBM, Microsoft, RSA, VeriSign)
The Web Services Security Policy Language defines a model and syntax to describe and communicate security policy assertions within the larger Policy Framework. It covers assertions for security tokens, data integrity, confidentiality, visibility, security headers and the age of a message.

*WS-Trust*
(being defined by IBM, BEA, Computer Associates, Layer 7 Technologies, Microsoft, Netegrity, Oblix, OpenNetwork Technologies, Ping Identity Corp, Reactivity, RSA Security, VeriSign, Westbridge Technology)
The Web Services Trust Language (WS-Trust) uses the secure messaging mechanisms of WS-Security to define additional primitives and extensions for security token exchange to enable the issuance and dissemination of credentials within different trust domains.

# 10. Bibliography and References

[Article1] http://www.nwfusion.com/columnists/2004/0524kobielus.html
[Article2] http://www.informit.com/articles/article.asp?p=169508&seqNum=1
[BRICKS04] Bricks deliverable – Aggregation Plan (D1.1.1)
[DNA5.1.2] eInfrastructure Reflection Group White Papers in conjunction with the EGEE Project Conferences, https://edms.cern.ch/document/545767
[EconomistSurvey] http://www.datasynapse.com/pdf/Economist_Intelligence_Unit_From_Grid_to_Great.pdf
[Economist, April 2 2005] "Google à la française", Mar 31st 2005, PARIS, From The Economist print edition
[eIRGWebSite] http://www.e-irg.org/about/
[eIRGWebSitedocs] http://www.e-irg.org/whitepapers/
[Fox99] Fox, E. A. Digital Library Definitions, 1999
[gLite] http://glite.web.cern.ch/glite/
[GGF] http://www.ggf.org/L_About/about.htm
[GridStart] http://www.Gridstart.org/download/BusinessNewsletterApr2004.pdf
[GridCafe] http://Gridcafe.web.cern.ch/Gridcafe/whatisGrid/reality.html
[IETF] http://www.ietf.org/overview.html
[IBMStandards] http://www-106.ibm.com/developerworks/views/Grid/standards.jsp
[IBMGrid] http://www-106.ibm.com/developerworks/Grid/newto/#5
[Johnsen] "Exploring Corporate Strategy, Fifth Edition", Gerry Johnsen and Kevin Scholes, Prentice Hall Europe, 1999
[Moore2005] R. Moore, A. Rajasekar, M. Wan, "Data grids, digital libraries, and persistent archives: an integrated approach to sharing, publishing, and archiving data", Proceedings of the IEEE, March 2005 (http://www.sdsc.edu/dice/Pubs/IEEE_Moore.doc)
[OASIS] http://www.oasis-open.org/who/
[OGSA] http://www.ggf.org/L_WG/News/OGSAFAQ_Handout.pdf
[W3C] http://www.w3.org/Consortium/
[WS-I] http://www.ws-i.org/about/Default.aspx
[WSDL] http://www.w3.org/TR/wsdl
[WSRF] http://www.globus.org/wsrf/
[WSRFStandards] http://www.globus.org/wsrf/specs/ws-wsrf.pdf
[WSRFGT4] http://gdp.globus.org/gt4-tutorial/multiplehtml/ch01s01.html
[WSRFFoster] "WS-Resource Framework: Globus Alliance Perspectives", Ian Foster, presentation
[SKYRME] The knowledge networking. Creating the collaborative enterprise. Butterworth-Heinemann. 1999
[SOAP] http://www.w3.org/TR/soap/
[SRB] SDSC Storage Resource Broker, http://www.sdsc.edu/srb/
[StanfordU99] Definition presented at the kick-off of the WG on Digital Library metrics, held on January 7th-8th 1998 at Stanford University.
[SunGridEngine] http://store.sun.com/CMTemplate/CEServlet?process=SunStore&cmdViewProduct_CP&catid=115672
[SunGridUtilities] http://www.sun.com/service/sunGrid/overview.html
[THREDDS] Thematic Realtime Environmental Distributed Data Services, http://www.unidata.ucar.edu/projects/THREDDS/Overview.html