SIXTH FRAMEWORK PROGRAMME
PRIORITY IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage

Information Society
Technologies

Sixth Framework Programme

Contract for:

NETWORK OF EXCELLENCE

Annex 1 - “Description of Work”

Network acronym: **DELOS**

Network full title: DELOS: a Network of Excellence on Digital Libraries

Proposal/Contract no.: **G038-507618**

Related to other Contract no.: IST-1999-12262

Date of preparation of Annex 1: **01 December 2003 (FINAL)**

Operative commencement date of contract: **01 January 2004**

1 Detailed joint programme of activities (JPA) - first 18 months

1.1 Introduction - general description and milestones

The DELOS network will carry out a broad range of interrelated activities whose combined effect should be to contribute to the successful achievement of the objectives as defined in Section 2. In particular, important objectives of the Joint Programme of Activities for the first months will be:

- *Tightening up the Network.* The first 18 month period is critical and much effort will be dedicated to constructing a sound community sense between the network members.
- *Putting up the infrastructure*. Because of its importance, we also plan to devote considerable energy to the building of an intranet portal, together with a publicly accessible web site, equipped with the appropriate services and communication tools for the all activities.

The plan for the Joint Programme of Activities defines the start and end of a subset of the activities presented in section 6. These activities are organized into clusters and are composed of three types of activities**: integrating, research** and **dissemination** activities. They have been assigned a schedule, both in terms of time and participating institutes and they favour the creation of the network community and emergence of multi-lateral collaborations. Each cluster is structured according to the Workpackage-Task hierarchy. In addition, there are the activities for the management of the Network, which also have been structured according to the Workpackage-Task hierarchy. The Network is thus composed of the following Workpackages for the first 18 months:

- WP1 Digital Library Architecture
- WP2 Information Access and Personalization
- WP3 Audio/Visual and Non-traditional Objects
- WP4 User Interfaces and Visualization
- WP5 Knowledge Extraction and Semantic Interoperability
- WP6 Preservation
- WP7 Evaluation
- WP8 Dissemination and Spreading of Excellence
- WP9 Assessment
- WP10 Administrative Management
- WP11 Scientific Management

In the following, the activities that the Network will carry out during the first 18 months are detailed for each Work Package.

1.1.1 WP1 - Digital Library Architecture

Information architectures for digital libraries are now evolving in parallel with architectures for peer-to-peer (P2P) systems, Grid-enabled environments and institutional repositories. In some cases, developments have proceeded in a fragmented way perhaps on a local basis (within an organization), in particular domains and sectors (within museums, libraries and archives) and within disciplinary boundaries (in bio-medicine or the performing arts). Synergies between initiatives are becoming apparent both at the technical level and in terms of the broad operating principles being adopted by the parties involved. For example, many of these developing architectures are predominantly service-oriented; they are adopting emerging Web Services standards and are becoming increasingly user-focused in their presentation. A main contribution of this workpackage will be to facilitate the development and integration of building blocks for digital libraries. This requires both the identification and specification of service interfaces and the definition of a generic digital library architecture that is highly customizable to individual application domains and national requirements. In the first 18 months we will concentrate on the following directions

- New approaches to the architecture for an "intelligent management" of digital libraries
- Enabling the coordinated Development of Information Architectures by an adoption of a set of common standards & protocols.
- Managing information dynamics and mobility

New Approaches to the "Intelligent Management" of Digital Libraries
We can distinguish three different approaches to DL architectures. The first is the (Web) service architecture (SA) and includes the related standards for describing, finding and invoking services. The SA leads to a new way of building distributed information systems, specifically DLs. Many applications do not need complete answers or fully fresh data. This brought up new distributed data management concepts subsumed under distributed Peer-to-Peer (P2P) data management. In parallel to this approaches a third direction, the Grid computing evolved and the associated Open Grid Service Architecture (OGSA)[1] is leading to the development of increasingly complex computer systems. These Grid architectures show high potential for the management, discovery, and load-balanced use of distributed digital library services, together with application-specific security aspects. Grid-enabled environments with distributed processing capabilities, the utilization of remote resources and dynamic online experimentation have encouraged the consideration of new approaches to information system management.

It is necessary to study these approaches in detail and to evaluate their advantages and disadvantages using adequate benchmarks that do not exist at present. Rather the fields are quite separated. This important step must be done during the first 18 months and will enable us to develop new intelligent architectures for a future DL architecture. The principles of autonomic computing[2], applied at the level of digital library applications and services, even extend the Grid architecture as a possible solution to addressing these challenges. This strand investigates the applicability of this approach to digital libraries through a mix of assessment and dissemination. Links may be possible to proposals following the Grid computing theme identified for the second call. The summary of outputs consists of:

- An evaluation of P2P, Grid and Service architectures to identify the benefits of each architecture for digital library applications. The evaluation includes the joint development of a benchmark and executions of the benchmark on selected different implementations of different architectures.
- Starting a joint demonstrator development integrating the benefits of the service architecture, Grid technology and P2P data management into digital library infrastructures.

Enabling the Co-ordinated Development of Information Architectures
This track describes the mechanisms, processes and tools that will be required to ensure the co-ordinated development of large-scale DL architectures. It is based on the principle of working towards common models, frameworks and platforms which span national boundaries, sectors and communities and which will be widely adopted and implemented. It sets out to build on significant existing work such as major national and international DL infrastructure initiatives, e.g. the UK JISC Information Environment[3], key standards developments e.g. Open Archives Initiative[4], and

[1] Open Grid Services Architecture http://www.globus.org/ogsa/

[2] IBM Autonomic Computing Manifesto
http://www.research.ibm.com/autonomic/manifesto/autonomic_computing.pdf

[3] The DNER Technical Architecture http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/dner-arch.html

emerging service models e.g. Web Services[5]. The outputs are designed to jointly facilitate the continued and sustainable development of a range of common models and frameworks.

These services will be based on established and emerging standards and protocols. The Web Services model is currently being promoted as a common foundation for e-service development however, a feasibility study is proposed to scope and assess the validity and durability of this approach and others, across a range of networked environments, sectors and disciplines. There is existing development of proof-of-concept demonstrators and pilot services to illustrate the potential of a set of shared infrastructure services. Examples include a service / collection description registry (e.g. JISC IESR), metadata schema registries (e.g. CORES, MEG), a cross-search broker service (e.g. Xgrain – EDINA), a resolver service (e.g. ZBLSA – EDINA), an ontology server (link to SEMKOS IP), and an authentication and authorization service (ATHENS), and a platform for the management of large-scale information spaces as e.g. ISIS/OSIRIS (ETHZ).

The summary of outputs includes:

- A comparison and feasibility study on the adoption of a set of common standards & protocols.
- Starting a joint development of infrastructures as demonstrator systems following selected standards & protocols

Managing Information Dynamics and Mobility

The combination of wireless and wired connectivity in a pervasive computing environment with increasingly small and powerful mobile devices, such as laptops, personal digital assistants, handheld PCs, and smart phones, enables a wide range of new digital library applications. We see an ever-increasing number of information providers and data sources, reaching from traditional databases and large document collections, information sources contained in web pages, down to information systems in mobile devices and embedded information in mobile “smart” objects. This leads us to considerably greater dynamics of information and as a consequence to the need for the infrastructure to keep track of dynamic information and mobility: information changes, information is replicated or is derived. New information providers and services appear any time. Clients connect and disconnect any time. After re-connection they want relevant refreshed information. The infrastructure envisaged here must be much more sophisticated compared to the state-of-the-art middleware in that it automatically performs – among many tasks - the following:

- It starts maintenance processes in order to keep replications and search engines consistent in case of new information or changes in the information.
- It keeps track of mobile clients, their location and their context and propagates relevant information after re-connections.
- In view of many concurrent processes the infrastructure executes them in a decentralized, peer-to-peer fashion and avoids central components as much as possible in order to be scalable and reliable.

This direction exhibits a close relationship to the WP on Information Access and Visualization, since it must provide the basic services for accessing and managing digital libraries via various mobile and non-mobile devices.

Summary of outputs:

[4] Open Archives Initiative http://www.openarchives.org/

[5] IBM Web Services Conceptual Architecture http://www-3.ibm.com/software/solutions/webservices/pdf/WSCA.pdf

- A quantitative evaluation of different concepts for synchronization and connection management and a synthesis of various approaches for the infrastructure of the dynamic digital library with mobile components
- Definition and implementation of basic services needed for supporting information dynamics and mobility
- A workshop on mobile information components for e-health monitoring as application of DL in Medicine

1.1.2 WP2 - Information Access and Personalization

The Information Access and Personalization (IAP) WP aims at the following strategic goals:

- Provision of a common foundation concerning the three major problem areas of work of the IAP cluster, namely “Information Access”, “Information Integration” and “Personalization”. Studies, workshops and other activities that lead to a uniform understanding of problems among researchers, will take place.
- Support of cooperation among individual IAP research groups will enable prototype systems and toolkits to be developed with the purpose of integration, improvement and evaluation of existing approaches.

To achieve those, the following activities are planned for the first 18 months. The main emphasis is on the tasks that will help establish a common basis on which future work will be based, as well as on issues that are already mature for being addressed.

Common foundation on Information Access: This task will attempt to obtain a common conceptual and infrastructure foundation on the topic of "Information Access". Specific activities in this direction aim at the collection and presentation of existing work on data and metadata models and query schemes for accessing homogeneous information stored in individual Digital Libraries.

Common foundation on Information Integration: This task will concentrate on problems that arise when one deals with multiple, heterogeneous digital libraries that need to be treated in a cohesive manner. Specific activities aim at the collection and presentation of existing work on data and metadata models and query schemes for access of heterogeneous information stored in distributed Digital Libraries.

A thematic workshop on the topics of the above two tasks combined (essentially dealing with information access in centralized and distributed, homogeneous and heterogeneous environments) will be held bringing together European and non-European researchers in order to survey the wide range of research efforts around the world. This will be a timely sequel to the two workshops on the corresponding topics that took place in late 2000 and the Fall 2001 under the aegis of FP5 DELOS NoE. The proceedings of this new workshop will be published on CD and the Web. The key findings of this activity will form the basis for a research agenda for various specialized topics.

Summary of outputs:

- “Information Integration and Access” Workshop

Common foundation on Personalization: This task deals with the topic of personalization and customization of the behavior of a digital library system. Specific activities include collection of information concerning user characteristics, study of user modeling approaches and possible forms of content and interaction personalization. A workshop on "Personalization" will help to collect and present of existing work in this field.

Summary of outputs:

- “Personalization” Workshop

Research on Information Access: In this task, new research will be conducted jointly by members of the cluster on problems that arise in the field of "Information Access" in digital libraries. In particular, the cluster will work on data and metadata models, query processing schemes and metadata generation algorithms for supporting information access in a single source. Different types of queries and models (taking into consideration multiple aspects of data, such as structure, semantics, context etc.) will be investigated and gradually supported in a modular query engine that will permit access of different types of data, namely structured, unstructured, and semi-structured. The first activity in this direction is to assemble a toolkit of a variety of algorithms that will operate under a unified framework.

Summary of outputs:

- Specification for toolkit of data searching algorithms

Research on Personalization: In this task, new research will be conducted jointly by members of the cluster on problems that arise in the field of "Personalization" in digital libraries. Different activities will take place. The IAP cluster will work on improvement and integration of specific models and algorithms for creating and maintaining user profiles and development and deployment of effective personalization approaches. A user profile should take into consideration different characteristics of the user, such as user behavior, preferences, and location. These characteristics are evolving and selection of the appropriate subset to be considered given a user request depends on a number of factors, including the nature and semantics of the request per se.). Privacy requirements (i.e., what a user is allowed to see) must be expressed and taken into account. Further research and developments will focus on models and techniques for support of personalization of user searches over multiple digital libraries. The ultimate goal is to construct a server for storage of profiles of users accessing one or more digital libraries and personalization of user searches on the basis of these profiles. The first activity in this direction is to assemble a toolkit of a variety of for user profiling and personalization of user requests that will operate under a unified framework.

Summary of outputs:

- Specification for toolkit of algorithms for user profiling and personalization of user requests

1.1.3 WP3 - Audio/Visual and Non-traditional Objects

During its first 18 months the Audio/Visual and Non-traditional Objects Cluster will pursue the following strategic goals

- Enable the integration of the different scientific communities that are involved in audiovisual content management in digital libraries and providing a common foundation and a common understanding of the developing state of the art in the field;
- Establish a shared understanding of the current generation of audio-visual digital library applications, testbeds and interface paradigms.
- Build on these first two activities to develop and integrate research results in new models for the extraction of audiovisual metadata, in audiovisual digital libraries access and retrieval models, and in models for the management of audiovisual metadata and content.

To achieve those, the following activities are planned for the first 18 months. The main emphasis is on the tasks that will help establish a common basis on which future work will be based, as well as on issues that are already mature for being addressed.

Joint activity on Audiovisual Libraries Forum

At the outset, a Forum on Audiovisual Libraries will be established to bring together information providers, digital library experts, and content experts with different scientific backgrounds like image processing, audio processing, video processing, pattern recognition, information systems, databases, and Artificial Intelligence. The objectives of the forum will be to establish communication and common understanding on the state of the art in the different fields, and the developments in the industry and in the audiovisual standardization bodies, the application trends,. The Forum will pinpoint significant trends and research directions in the field and develop a a road-map for achieving the ten year mission statement for the digital libraries in the audio visual areas. The Cluster will establish a reference database to point to state of the art audiovisual prototypes, applications and interfaces, as well as multi-channel delivery of audiovisual digital library content, including content outside Delos. This will deliver key information about state of the art implementations useful to the industry and researchers. In parallel, public access, domain specific audiovisual content will be generated and/or assembled, in order to create possibilities for experimentation with extraction, retrieval and management algorithms.

Summary of outputs:

- Workshop on past and future of audio visual digital libraries and their applications

Common Foundation and Joint Research on Metadata Capturing for Audio-Visual content

This task focus on solutions for the extraction of content metadata and their integration with standard content description frameworks for the purpose of effective retrieval. Particularly, the task will focus on automatic and semi-automatic information extraction models and suitable intelligent interfaces to facilitate the integration of automatically extracted knowledge with user-provided data. It will also address domain-specific methodologies for automatic extraction of content descriptions at the semantic level. The task will cover the analysis and verification of the state of the art and the study and development of new prototype solutions that integrate and evolve previous achievements

Summary of outputs:

- Report on the state of the art of metadata extraction for audio-visual DLs
- Toolkit of algorithms for metadata extraction

Common Foundation and Joint Research on Universal Access and Interaction with Audio-Visual Libraries

This task deals with the problem of accessing content of audio-visual digital libraries through different access terminals and delivery media and issues related to user interaction and dialogue with the system. It will cover solutions for content based retrieval and paradigms and interfaces to interact with the content, user’s cognitive models to interact with multimedia content, and methods to deliver content according to both the user’s preferences and cultural profile, and the user’s terminal characteristics. The task will also investigate the integration of preexisting and emerging content specifications so as to develop efficient personalized added value services on existing broadcasting channels. The task will include the analysis of the state of the art and the development of prototype systems of multimodal interfaces, content retrieval, semantic transcoding and terminal adaptation.

Summary of outputs:

- Report on the state of the art on Universal Access and Interaction for audio-visual DLs
- Toolkit of algorithms for universal access and interaction

Common Foundation and Joint Research on Management of Audio Visual Content in Digital Libraries

This task will address integrated analysis and research development in the management of audiovisual metadata. It will focus on new approaches for audiovisual summarization, on the management of domain specific and context specific metadata, and on models. It will also cover methodologies for providing interoperability across standards in order to map metadata which follow different audio-visual standards. Finally, the task will address the problem of defining models of user categories that capture the stereotypical way in which they work with multimedia data, and build effective tools for filtering data and acquire from the library the appropriate knowledge to the user’s profile.

Summary of outputs:

- Models for interoperable domain-specific metadata following different standards.
- Toolkit of algorithms for user profiling and personalization of user requests, data filtering and summarization

Joint Activity on Demonstrators and Testbeds
In this task the cluster will assemble and/or reference a number of tools and solutions that have been proven affordable and interesting so that they can be available to the general community and give hints for future applications and developments. They will also demonstrate the practical feasibility of advanced applications for audio-visual libraries, Testbeds will be made available to researchers and developers that include image, video, 3D graphics, audio, music, and more in general multimedia objects that can be used as a reference set to test and compare the performance of different solutions. Testbeds will include also currently available data sets for audiovisual digital library applications.

Summary of outputs:

- Web site with links to demonstrators and test datasets for audio visual digital library applications
- Workshop on audiovisual digital libraries (achievements foreseen applications)

1.1.4 WP4 - User Interfaces and Visualization

The major goals of the User Interface cluster during the first 18 months period are the exploration of the role and scope of DL user interface research, the investigation of the requirements for a DL interface design that accords support to the user throughout the entire DL lifecycle, and the development of a generic theoretical framework for DL interface design. It is worth mentioning that the DL arena is presently characterised by ad-hoc user interface designs.

From a more global perspective, various thematic workshops will be organized and a shared workspace set up to collect inputs and surveys from DELOS projects and other existing projects. Eventually, but not necessarily in the first 18 months, the cluster will edit a book collecting the most notable DL contributions.

Provision of an Empirical Basis: Towards realising the aforementioned User Interface cluster goals, the cluster will initially carry out various studies in order to realise an empirical basis for the research. These studies will focus on the degree to which the design of interface metaphors are based on empirical identification of user requirements, of explicit descriptions of the (collaborative) tasks situations that DL will support as well as on empirical studies of the characteristics of different work domains (e.g. virtual organizations, libraries, archives, schools, industry, etc.) The data for this overview and the different analyses will be based on ongoing DL research projects and the DL literature. The result of this analysis will be an empirical information base that can provide a common ground and a basic insight that can be used by all the participants in the other tasks of this cluster. Similarly, other cluster activities will provide significant input to the further development of the information base into a taxonomy of interface metaphors. A taxonomy based on the collaborative effort and the comprehensive expertise in the cluster will provide a significant conceptual contribution to the teaching and development of the next generation of DL interfaces.

DL Lifecycle Support: The User Interface cluster too will investigate the phases of which the DL lifecycle is made. The cluster will then identify user needs/functional aspects pertaining to DLs. The DL lifecycle phases will then be mapped to the DL functional requirements. This cluster will establish a close cooperation with libraries in order to have a more solid and first-hand understanding of their operations/functionality. Moreover, non-functional requirements (e.g., adaptability, scalability, platform independence, abstraction, etc), which are seldom addressed, will be analysed. Note that these requirements are becoming increasingly critical especially for DL systems which are intended to have long life cycles, an orientation towards the citizen (including young, elderly, people with disabilities, users with variable interests, requirements or competencies).

Characterisation of DL Users: The cluster will carry out a characterisation of DL users and user communities. A classification scheme will be developed through which user types and tasks will be classified. The cluster will also come up with a formal descriptive framework for user types. In order to realise the foregoing, a questionnaire will be set up on the Web through which DL users may give in their input. Based on the understanding that the research in DLs has largely neglected librarians, this cluster will exploit the aforementioned close cooperation with libraries as an avenue for interacting with the librarians. The cluster will also work closely with governmental departments, organisations, and societies that look into the plight of the disabled. This is intended to help in learning the special needs of the disabled/impaired users.

Context Consideration and Exploitation: The physical, organizational, social, and technical environments/context of the DL user will be analysed and exploited in order to meet his/her needs. In particular, the User Interface cluster will study context models. It will then develop a formal context description and a specification language. The foregoing realizations will then be exploited to model context-dependent DL functionalities. The cluster will exploit pertinent research avenues such as workshops and conferences in arenas such as ubiquitous computing, wireless technologies, and mobile devices. The exploitation will involve studying of, participating in and contributing to such forums.

Visualisation in DL Systems: DLs have not yet fully exploited what the field of information visualisation has already delivered or can deliver. The User Interface cluster will investigate the exploitation of relevant existing and/or novel visualisations in presenting DL results/views. To realize that, the cluster will explore various visualisation techniques and efforts, and how they are or can be used in DLs. The cluster will also develop appropriate novel DL visualisations. All the relevant existing DL visualisations and new DL visualisations will be collected and arranged in the form of a DL visualisation suite. In cooperation with the Evaluation cluster, the various visualisations in the DL visualisation suite will be subjected to usability tests. The issues of effectiveness, expressiveness and interactivity of the DL visualisations will be especially crucial to investigate.

Theoretical framework for DL User Interface Design: The User Interface cluster will also build a comprehensive theoretical framework from which DL user interface designers/engineers can develop effective DL user interfaces. The User Interface cluster will work closely with the other DELOS partners in order to acquire input such as scenarios, requirements, interface designs, and evaluation results. Each input will be reviewed from a more general perspective than the one defined by the specific DELOS partner or project providing the input. The inputs will be used to identify patterns in DL user requirements/functional needs. The cluster will then link the user requirement patterns to relevant DL design representations/metaphors. The User Interface cluster will also elaborate formal models that map DL user requirement elements to DL design elements.

1.1.5 WP5 - Knowledge Extraction & Semantic Interoperability

The Knowledge Extraction & Semantic Interoperability WP has two key strategic goals:

- To co-ordinate a programme of activities which brings together research excellence from a range of inter-related knowledge engineering and information management areas, and which facilitates the sharing of experience and expertise amongst practitioners from both DL and Grid/computing science backgrounds.
- To explore the potential of new models, algorithms, methodologies and processes in a variety of technical applications, institutional frameworks and cross-sectoral environments, which will lead to the creation of guidelines and recommendations of best practice for dissemination to the widest possible community of interest.

Research on Information Repositories: These activities will together provide a structured framework for the integration and dissemination of WP outputs, opportunities for researchers in the field to meet and discuss their findings and for the exchange of expertise and understanding with the wider community.

Research on Knowledge Extraction & Semantic Interoperability: The research tasks are designed to initiate interaction between institutions and individuals and to provide a sound basis for further work in the future. They have been selected as items of high interest and debate currently amongst the global digital library community, but which will also bring together researchers from different disciplines and backgrounds (e.g. DL and Grid). These activities will also involve members of academic faculty who are frequently the creators of data and information which is re-purposed in the scholarly knowledge cycle, and who have an important stake in maintaining the integrity and quality of resources available to their colleagues and peers.

The research will break new ground by interlinking ideas and practice from areas which are currently better documented and understood by the DL community with new associated applications which have not yet been considered in great detail (e.g. institutional e-print repositories which are being widely implemented in comparison to e-data repositories).

Ultimately the outputs of the research activities will enable the creation of briefing papers, guidelines and best practice recommendations which will inform practitioners and managers within institutions who are engaged in service delivery to the community.

1.1.6 WP6 - Preservation

Research in the area of digital preservation is fragmented and in need of integration. From the array of possible research tasks, the Preservation Cluster will focus on those designed to initiate collaborative interaction between institutions and individuals, focus and enable digital preservation, and deliver tangible results by bringing together fragmented research results in different laboratories. The Preservation Cluster has four strategic goals:

- To eliminate the duplication of effort between research activities by creating an integrating framework to co-ordinate and promote research and projects and to enable identification, collection, and sharing of knowledge and expertise;
- To examine core issues that will deliver essential guidelines, methods, and tools to enable the construction of preservation functionality within digital library activities and deliverables are created.
- The establishing of testbeds and validation metrics. These will provide a framework for testing preservation strategies, for establishing the preservation worthiness of digital library implementations, and create greater comparability between research and implementation activities.
- To relate the digital preservation research agenda more directly to the development of exploitable product opportunities and to develop links with the industrial sectors.

To achieve the first of these goals the preservation cluster will promote work that establishes frameworks, metrics, validation sequences and extensible and useable typologies.

There are a lack of digital preservation testbeds and this reflects a lack of agreed metrics for establishing distributed testbed environment in this arena. If we are to ensure the success of research in areas such as digital repositories environments where different solutions can be evaluated provide a critical step. The cluster aims to produce, test and disseminate testbed design, validation, comparability metrics and a test data set. The results should be useful for testing and validating digital preservation methods, repository implementations and creating metrics for comparability between testbed environments. These developments will build on the experience within the research community. As a first step in this direction the cluster will review and establish metrics for testing and validating digital preservation strategies.

Digital repositories have a central role in the long term curation of digital objects and models have begun to emerge as to how these repositories should be developed and deployed. So far the effectiveness and viability of repositories have not been evaluated. Beyond the political, economic, and organizational issues research into the technical issues remain challenging. Problems such as addressing the complexity of engineering generic connections to enable newer hardware to communicate with legacy peripheral devices and the definition of automated testing sequences to enable cost effective assessments to be made of the software in the repository to determine whether it continues to function and behave as it was originally designed to do need to be addressed. A first key step is ensuring access to tools to assess how effectively and efficiently digital library implementations provide preservation functionality. This activity will focus on defining frameworks and mechanisms to audit and certify repositories and therefore to assess whether or not Digital Library implementations achieve any measure of sustainability. In this area the cluster will contribute to the development of digital repository frameworks and the development of methods for validating the suitability of digital repository implementations. An initial output of this work will be a report on technical aspects of digital repository designs and storage models.

Researchers do not know enough about the relationship between file formats and digital preservation, although it is agreed that the properties (or attributes) of some formats may pose a greater or lesser challenge to their preservation and that different formats require different approaches to modeling the preservation activity. Of course in the archival and digital library environment multiple file formats need to be managed simultaneously and they have different rates of obsolescence. Until very recently accessible sources of information about digital file formats (e.g. syntactical and semantic form) have not been the subject of systematic, consistent, comprehensive, and quality assured data collection practices. Digital Library initiatives will be able to use this information most effectively if the formats themselves can be handled as classes. The cluster will collaborate with other international activities to ensure the development of open source file format registries and the creation of processes to make these registries useable as tools to enable the preservation process. Information about file formats gain most value when they can be related to preservation methods and typological frameworks. Establishing a typological framework for file formats that is extensible enough to incorporate new formats as they emerge is a critical step. During the first phase of the NOE the cluster will deliver a report on file format registries and the relationship between them and preservation strategies.

Current research has not succeeded in developing approaches to functionality and behaviour abstraction and representation. Efforts to carry this work out depends on defining key attributes, establishing a representation framework, and engaging researchers in formalisation methodologies and theory. In the first of a two stage process the Preservation Cluster will establish a workgroup to define what kinds of functionality and behaviour metrics are required if digital libraries are to be able to verify automatically whether or not system behaviour and functionality match that which the application had originally following its migration, emulation, or retargeting. This will enable us to handle the creation of representations that can be used to establish benchmarks to measure consistency of performance across migrations or emulations. As a first step to ensure the viability of metrics the attributes of functionality and behaviour need to be defined and mechanism for representing them established. In second step mechanisms for validating them need to be generated. The cluster will produce a study on functionality and behaviour attributes and verification metrics and a mechanism for representing them.

For some types of digital libraries it may be feasible to build preservation functionality into systems themselves, this is especially the case for those designed with a direct relationship to systems creating the digital objects they will contain, but for most the systems need to cope with the ingest of digital objects in different formats. This means improving our knowledge about what preservation functionality really is and ensuring that this functionality can be effectively communicated to system developers, modelled and implemented by them. The cluster will define the requirements for a preservation modeling tool that can be used alongside system design and development methodologies to address preservation issues in system analysis and design activities. The results of this work will be presented as a suite of guidelines and methods for analysing and defining requirements for preservation functionality within widely used system development methodologies.

While digital preservation has attracted substantial attention within memory institutions, it still has not attracted sufficient numbers of academic and industrial researchers. Among the reasons for this are the limited funding opportunities and the lack of understanding of the digital preservation issues and the research opportunities. While the Preservation Cluster focuses on synergizing research it also aims to contribute to the development of new funding opportunities, the establishing of learning opportunities for younger researchers, and the creation of an online learning environment that could be used to develop expertise among existing and potential researchers. The cluster will work to raise awareness among national funding bodies of the problem and the range of research activities that need to be supported at national level to foster research by unlocking the purse strings of funding agencies. It will develop a framework for a digital preservation summer school as an annual event to introduce between thirty and fifty younger Digital Library researchers to digital preservation researcher issues and opportunities. It will create a framework for a Digital Preservation Educators Forum to deliver online learning in current preservation research to students, researchers and professionals.

1.1.7 WP7 - Evaluation

The evaluation WP aims at two strategic goals:

- Supporting current activities in the area of DL evaluation by establishing an evaluation forum which collects existing toolkits and testbeds, and provides mechanisms for communication between DL developers and evaluators. In addition, current evaluation initiatives which are relevant for DL-related activities will be continued and extended.
- Performing research on evaluation methodology by formulating a conceptual DL model, developing new evaluation approaches on the basis of this model, specifying corresponding methods for these approaches, implementing toolkits to apply these methods and developing testbeds for specific types of content and usage

DL Evaluation Forum: In order to achieve the first goal, an evaluation forum will be created; the INEX and CLEF evaluation initiatives will be continued and extended within this framework.

The evaluation forum will collect existing methodologies, testbeds and toolkits for DL-related evaluation. For this purpose, workshops and working groups will be established, in order to acquire a survey of current practices. In addition, the metalibrary of test collections acquired by DELOS under FP5 will be extended: the testbeds gathered so far are mainly collections, more attention will be given to testbeds that also include management as well as usage components. These testbeds will be made available in electronic form, in order to disseminate knowledge on DL evaluation tools and to provide research groups with easy access to existing testbeds. A second important element of the evaluation forum will be the support for communication about evaluation issues, especially between evaluation specialists and DL developers. An electronic discussion forum will be set up for this purpose. By bringing together researchers from these two areas, bilateral prototype evaluations can be arranged, where an evaluation specialist would work together with a developer needing evaluation for a new type of functionality.

Whereas typical documents in DLs have an explicit internal structure, most current evaluation initiatives only deal with unstructured documents. For this reason, the INEX initiative for the evaluation of information retrieval on XML documents was started in 2002, with 35 active groups (40 groups have registered so far for 2003). Current usage types are ad hoc-queries in two variants, namely content-only queries searching for relevant portions of documents, and content-and-structure queries combining content and structural conditions. Retrieval effectiveness and efficiency are the evaluation criteria currently considered. Appropriate evaluation measures that consider structural relationships between different answers are still to be developed. Other usage types will be examined in the future, especially interactive retrieval involving browsing, navigation and iterative retrieval. As INEX is currently restricted to text-only documents, the initiative will also develop new testbeds for structured multimedia documents, such as MPEG7 annotated documents.

Cross-language information retrieval (CLIR) is a key area in the digital library domain, especially in Europe where so many of the collections consist of material in multiple languages. Retrieval systems must thus be able to efficiently cross both language and cultural boundaries. The Cross-Language Evaluation Forum (CLEF) has created an infrastructure and organized campaigns for the evaluation of mono- and cross-language information retrieval systems since 2000. So far, CLEF has focused mainly on testing text retrieval CLIR system performance. The CLEF activity within DELOS under FP6 will be specifically aimed at evaluating components of multilingual retrieval systems that are of particular relevance to the digital library application area. The activity will thus concentrate on three main issues in the first 18 months: user needs, cross-language retrieval on multi-media documents, evaluation of multilingual retrieval systems on the web. The user needs area will address two questions: the testing of systems designed to assist users searching for relevant information in unfamiliar languages with query formulation or results interpretation (interactive systems); the testing of systems that retrieve exact answers rather than entire documents in response to specific queries (question answering systems). With respect to multimedia, the activity will test cross-language systems running on digital audiovisual archives, a relevant sector of the digital library landscape. Cross-language spoken document retrieval (CL-SDR) extends the IR framework by assuming not only that queries and target documents are not in the same language, but that the target documents are in spoken form, e.g. recordings of broadcast news. In consideration of the enormous increase of non-English material on the world wide web, another important objective of this task will be to study a methodology and create an appropriate test-bed for the evaluation of multilingual web search engines.

Research on Evaluation Methodology: This activity will start at a rather theoretical level with the development of models and methods, followed by the implementation of appropriate toolkits and testbeds. As a first step, a conceptual DL model will be developed, which describes the structure and behavior of DL systems. Major building blocks for such a model will be classification schemes for content, management and usage. The model will form the basis for relating the evaluation of different aspects of DLs to one another.

Standard evaluation methods will be specified, starting from this model. For this purpose, a meta-analysis of existing studies will be performed, along with collaborative efforts to compare and evaluate techniques for DL evaluation. On the basis of the results of these activities, standard techniques, methods and measures can be developed and be made available to the entire DL community. In order to ease application of these evaluation methods, appropriate toolkits will be developed, especially for user studies and user-centered evaluation. The tools to be provided should support easy prototyping of interfaces, gathering of user input (logs) and data, and analysis of user input. Ultimately, these tools will support the creation of a simulated usability laboratory. Since there is a lack of testbeds for usage-oriented evaluation, specific effort will be given to the creation of appropriate testbeds. For this purpose, a database containing user interaction data collected from existing DLs will be created. As a second approach for a usage-oriented testbed, a framework will be set up, comprising both content and management components (a modular DL system), where research groups can plug in their own services for evaluation purposes.

1.1.8 WP8 - Dissemination and Spreading of Excellence

A detailed description of the different types of Spreading-of-Excellence activities was provided in Section 6.3. The Dissemination Workpackage will rely on the support of a Virtual D-Lib Competence Center, which will be constituted by three physical sites: CNR-ISTI, UKOLN (University of Bath) and NetLab (University of Lund). In the first 18 months a great emphasis will be placed in the establishment of a web site, equipped with the appropriate services and communication tools for the all activities, in addition to the organization of scientific workshops and awareness events. The details of those activities are provided in Section 9.6.8.

1.1.9 WP9 - Assessment

This Workpackage has two specific objectives:

- to review and assess the overall activity of the Network and to provide recommendations

- to monitor and facilitate the integration process of the Network

The first task will be to establish an Advisory Board of external experts which will be responsible for a periodic assessment of the activities of the Network, as described in detail in Section 8.

The second task will cover the integration of the different activities that are undertaken in the different work packages within the first 18 months of the project and, if needed, for the following period. According to the integration activities that have been discussed in Section 6.1, integration will cover:

- Definition and development of tools and solutions to enhance integration and information exchange between clusters
- Definition and activation of discussion forums and events for information exchange between clusters
- Definition and implementation of standards and interchange formats of the materials developed by the different clusters

For the carrying out of these activities the project will establish an “Integration Task Force” composed by one leading researcher per cluster, which will be responsible for defining the strategies to be followed for these three activities. It will closely cooperate with the Scientific Board, reporting on the status of integration, identifying critical aspects and possibly suggesting solutions.

The Integration Task Force will coordinate its activity closely with the Workpackage Leaders.

1.1.10 WP10 – Scientific Coordination

The scientific coordination workpackage will direct and supervise the scientific work of the Network. In particular, it will (i) organize the Network as a whole, (ii) supervise the scientific progress of the Network, (iii) ensure that all deliverables are available on time, (iv) create and maintain the conditions necessary for successful collaboration, (v) represent the Network in concertation with other scientific events.

The Network will adopt the following coordination structure:

- Advisor Board (AB)
- Scientific Board (SB)
- Workpackage Steering Committees (WPSCs)
- Virtual D-Lib Center Management Committee
- Members General Assembly (MGA)
- Specific Task Forces (STF)
- Scientific Coordinator
- Administrative Coordinator

1.1.11 WP11 – Administrative and Financial Management

The administrative and financial management workpackage will ensure a strong and coherent administrative and financial management of the Network. The administrative and financial coordinator will also handle the reimbursement of the expenses incurred by the (invited) participants to events and meetings organized by WP8, WP9 and WP10 (e.g. workshops, national events, Advisory Board and Scientific Board meetings, etc.). For this reason, a large part of the money budgeted for those activities has been allocated to WP11 rather than to the WPs responsible for the events.

This Workpackage will, in particular, (i) handle all the administrative tasks connected with the NoE's activities, (ii) handle all the financial tasks connected with the NoE's activities, (iii) provide a global Intellectual Property Rights (IPR) frame for the whole participants, (iv) ensure institutional exchanges with the EC representatives, and (vi) promote the gender equality within the Network.