The DELOS Network of Excellence on Digital Libraries[1]

Recommendations and Observations for a European Digital Library (EDL)

In response to the European Commission's call for online consultation

[1] Funded by the Information Society Technologies (IST) Programme of the European Commission (Contract G03-507618)

Executive Summary

On December 5 - 6, 2005, the DELOS Network of Excellence on Digital Libraries held a brainstorming meeting in Nice, France, to formulate responses to the i2010 Digital Library questions from the European Commission. We fully support a **European Digital Library (EDL)** that will achieve a more inclusive European Information Society and overcome the geographical and social digital divide and recommend **a two-pronged strategy:**

1. **Put a short-term focus on extensive digitization of materials and their organization into a system with unified access across Europe**, as envisioned in the i2010 Digital Library initiative, using technology that anticipates the long-term vision.
2. Build on this infrastructure to pursue the **long-term vision of a much broader system of highly interlinked information and services that will provide very rich functionality, supporting new ways of intellectual work, communication, and process execution in business, government, and daily work.**

The first set of recommendations **responds to the Commission questions**:

- Perform **incremental user requirements analysis** to support decisions at every step of the development of an EDL so that the result are optimized to meet these requirements.
- **Coordinate digitization efforts** among stakeholders through a coordination infrastructure to avoid duplication of effort. **Encourage creation of borne-digital material** in the proper document format right from the beginning to reduce the need for digitization in the future.
- Develop comprehensive **metadata registries** to support access to and interpretation of the digitized or born-digital material. Follow standard metadata schemes whenever possible.
- Give particular emphasis to digitization and management of **multilingual material and of multimedia material**. Develop lexical and ontological resources that allow for mapping between European languages. Develop tools for feature extraction and automatic metadata creation for multimedia data.
- **Deploy advanced search engines** that improve access to digital material now and **allow for integration with other components in the future**, so that entire scientific or business workflows can be implemented to leverage information for increased productivity.
- Establish **the appropriate legal and organizational framework to streamline the process of obtaining permission for the use of copyrighted material under diverse circumstances**, including when the rightful owner is unknown.
- Base the EDL on a **service-oriented architecture that provides users with seamless access to fee-based information** from publishers and other providers.
- Encourage **preservation** to ensure the long-term viability of the European Digital Library:
  - Establish a clear **legal framework** that mandates ownership of digital objects to the producer as an incentive for preservation.
  - Introduce **legislation so that, within the EU, deposition of an object to a repository in a single European country is sufficient.**
  - **Conform to international standards,** especially for unique object identifiers.
  - Capture **provenance information** among metadata to ensure traceability.
  - Encourage **community consortia that will ensure the maintenance of multiple distributed copies** of each object as a protection against physical loss.

- Establish **service centers for monitoring technology obsolescence** to inform organizations of obsolescence risks, provide migration services, develop emulation methods, and preserve obsolete technology for as long as possible.

Second, we recommend several **immediate actions whose results would be very beneficial if they are available before large-scale digitization or other processing starts**.

- To **bridge the current gap between research and application**, commission the preparation of several practice-focused, solid, and exhaustive **state-of-the-art reports and tool repositories** and establish several **competence centers** that are accessible through a one-stop portal. **Such reports are needed especially in the following areas**:
  - tools for digitization, OCR, and other related technologies;
  - metadata standards for all kinds of objects;
  - multilinguality, multimedia, and other complex objects;
  - search engines, annotation, personalization, collaboration, data integration;
  - preservation issues, strategies, and techniques.
- Establish large-scale **test beds** for the speedy evaluation of research and development results to **shorten the time from research to application**.

Third, we propose **general principles** for the development of a European Digital Library:

- Use a **user-centered and task-centered** approach to design and development of the EDL.
- Support **unified access to materials from libraries, archives, museums,** other memory institutions, and ultimately, the Web.
- Configure the EDL as a **federation of interoperable components**
- Use a **unified service-oriented framework for all digital library projects sponsored by the European Commission** as part of a single European Digital Library system spanning several subject domains.
- Fund **research and development that is focused on enabling the long-term vision**.

Forth, we present the **grand DELOS vision for Digital Libraries:** Digital libraries will become the **universal knowledge repositories and communication conduits for the future**, common vehicles by which everyone will access, analyze, evaluate, enhance, and exchange all forms of information. Digital library systems will

- be person-centric,
- support user-to-user communication and collaboration,
- operate in a globally distributed environment, and will
- serve “all” applications and “all” forms of content.

Reaching this vision requires advances in many technological areas, among them:

- providing functionality that **supports users’ work and daily activities**;
- managing **ontologies as core components of digital libraries**, for collaboration, large-scale integration, and other functionality;
- employing rich user profiles to **personalize** the behavior of the system at several levels;
- supporting **pervasiveness of information**, mobility of users, variation in content quality, and dynamicity of access devices through location and context awareness; and
- **processing incoming data streams** through aggregation and integration.

Introduction

On December 5 - 6, 2005, the DELOS Network of Excellence on Digital Libraries held a brainstorming meeting in Nice, France, to formulate responses to the i2010 Digital Library questions from the European Commission and to discuss the DELOS vision of the future of Digital Libraries. Participants came from within and without DELOS, from Europe and the United States, including librarians, researchers in the DL field, and representatives of the European Commission (see Appendix 1 for a list). This report distills the results of the meeting.

At the meeting, much discussion focused on the requirements for a future European Digital Library (EDL) taking into consideration existing efforts such as The European Library (TEL) and the MICHAEL initiative. There was consensus on the significance of large-scale digitization of materials and their organization into a system with unified access across Europe, as envisioned in the i2010 Digital Library initiative. There was also consensus that such a European Digital Library can and should develop in the future into a much broader and richer system of highly interlinked information and services, created by collaborating users, and augmented by inputs from automatic data capture, in which the boundaries between reading, annotating, and authoring become fluid and which will provide new and rich functionality, some explored by DELOS and other research efforts and some not yet imagined. Such a system will support new ways of intellectual work, of communication, and of executing processes of business, government, and daily work. There is the potential of creating a true European Information Space. The potential exists for digital libraries to become the universal knowledge repositories and communication conduits for the future, a common vehicle by which everyone will access, discuss, evaluate, and enhance information of all forms.

The report is organized as follows:

Section 1. **Responses to the European Commission questions**

Section 2. **Short-term actions**
Makes recommendations for short-term activities to prepare for efficient and effective implementation of a European Digital Library

Section 3. **General principles for the development of a European Digital Library**
Makes recommendations that should contribute greatly to the short-term and long-term success of a European Digital Library

Section 4. **Vision for the future**

Appendix 1. **List of participants**

Appendix 2. **Faceted classification for the organization of issues arising in the European Digital Library initiative**

Appendix 3. **Inventory of additional ideas from the DELOS brainstorming on an EDL**
Contains much useful detail for both the short term and the long term

.Appendix 4. **Selected list of relevant publications from the DELOS community**

1 Responses to the European Commission questions

The responses given in this section focus on the following:

- Engaging private companies in the work; therefore, issues related to stimulation of private-public collaboration underpin several of the responses
- A service infrastructure framework, which will allow collaboration among libraries and among libraries and private enterprises, offering either information services, payment services, or other relevant library services
- Multilingual access, as this is considered one of the most important aspects to ensure the survival of the multicultural dimensions of the European Cultural Heritage
- Access methodologies, which will use the detailed and structured distributed information available and will adapt to both the working conditions of the searcher and the changing mechanisms of access (e.g., pda, etc.)
- Preservation activities

Digitization and online accessibility

Question 1

What additional measures could be taken at national and European level to encourage digitization and online accessibility of material in all European languages?

Recommendation 1.1[2]

Assemble data about the needs for digitized materials and about potential users' willingness to expend resources (time and money) to gain access to digitized materials. Use this data in planning digitization efforts.

- Monitor, collect, and evaluate user requests to libraries, information services, and on-line catalogues. Build demand monitoring into access systems.
- Assemble data from past and future studies on digitized material needs.
- Elicit industrial research needs.

Recommendation 1.2

Aim at a collaborative digitization effort that avoids duplication. To make this possible, provide a web-based infrastructure to keep track of who did digitize what and to plan who should digitize what, including digitization on demand.

The collaboration should be worldwide (e.g., Google or the Library of Congress "World Digital Library")

Day-to-day coordination on who should digitize what is essential for the optimal use of resources. This needs good technical infrastructure:

[2] Recommendations 1.1 – 1.3 are especially urgent since they would greatly improve the efficiency of digitization effort and the interoperability of the results.

- A federated union catalog of participating libraries (building on the federated union catalog of national libraries established by TEL). This would serve as the foundation of a Web-based system for checking for works already digitized or in the process of digitization (completed, in process, planned), and for recording works as they are digitized (completed, in process, planned).
- Facilitation of online consultation as to which of several libraries owning a work will digitize it.
- Support for digitization on demand: Interlibrary loan requests would be met by digitizing and making the material available in digitized form. For copyrighted materials, this may entail a user fee. The user should have the choice of obtaining a hard copy as in traditional interlibrary loan.

Recommendation 1.3

Building on the efforts of TEL and other existing metadata registries, develop a comprehensive metadata registry (possibly distributed) for the European Digital Library, giving detailed information on metadata schemes for all types of materials.

- The European Library is creating a European Metadata Registry that will hold the following:
  - the TEL Application Profile (an XML name space specifying the metadata elements used in TEL)
  - crosswalks from the metadata schemas of the participating national libraries and from major metadata schemes such as USMARC21 to the TEL Application Profile
- Work towards an ontology-driven (semi-) automatic mapping among these schemes, identifying agreements and differences. This will facilitate the automated import and export of schemas.

Recommendation 1.4

Building on widely accepted frameworks, such as the Dublin Core and OAI, fashion as much agreement as feasible on metadata schemes used both within each type of material and across types, including metadata on the digitization process itself.

- This needs to take account of the metadata already existing for collections being digitized in order to minimize costs. Automated mapping of metadata following an existing scheme to a new, standard scheme, using the information in the metadata registry, will increase efficiency.
- Develop common reference models for metadata with the same purpose and application area

Recommendation 1.5

For digitization as well as access support, develop a multilingual component that pervades all other system components.

Recommendation 1.5.1

Provide a guide to efficient localization tools for interface developers, to allow for low-cost adaptation of interfaces to many languages.

Recommendation 1.5.2

Develop lexical and ontological resources that allow for mapping between European languages. Start with providing multilingual access in TEL, through mapping controlled vocabularies in several languages, with the subject vocabularies of the national libraries being the first candidate.

In the long run, this should be developed into a comprehensive set of interrelated lexical and ontological resources that allow for accurate mapping between all European languages in all subjects.

Support for information processing and searching in multiple languages requires Knowledge Organization Systems (KOSs) that can bridge between languages, specifically for crosswalks among European national libraries' vocabularies. The initial aim should be an 80 - 20 solution: rather than aiming for perfection with a very large effort, one might be able to achieve 80% of the result with 20% of the effort. Several components are needed:

- a multilingual subject KOS - initially it should be a KOS that mediates between the subject heading systems of the national libraries in participating countries
- a multilingual gazetteer
- a multilingual authority list of names for persons and organizations

These tools should be developed by a distributed, collaborative effort within the EDL framework, building on and integrating the many existing multilingual KOS such as the European Union's multilingual vocabulary and the CENL (Conference of European National Librarians) projects MACS (Multilingual access to subjects) and MSAC (Multilingual Subject Access). Development should include semi-automated methods for matching across languages, using all available vocabularies as well as methods for adding vocabularies in other languages (a short-term R&D problem). Another very useful method for detecting term mappings is mining cataloging records for the same book in several of the participating national libraries (method building on work done by Michel Buckland in Berkeley).

Recommendation 1.6

Put more focus on digitization of and access to multi-media content.

This requires decisions on document formats and additional functions of search engines.

- *Automatic annotation of multimedia data*
  Existing multimedia retrieval systems either rely on manually generated metadata or provide users with similarity search capabilities. Manual generation has a high cost, while state-of-the-art similarity search techniques still suffer from a "semantic gap": They retrieve physically-similar rather than semantically-similar documents. Advanced techniques for automatic generation of metadata for multimedia documents should be developed. Generated metadata should include descriptors that allow semantic similarity search and should provide semantic content annotations as well. Innovative research is also needed in the field of automatic extraction of rhetoric, affective, and emotional descriptions.
- *Efficient retrieval and access to multimedia data*
  Efficient and effective retrieval of textual data is supported by existing and well-proven techniques, widely used in commercial search engines, but there is a lack of access methods that offer comparable performance on multimedia data. Innovative research to enhance present state-of-the-art techniques in order to provide very fast response to multimedia queries, under high data and usage loads, should be developed.

Research directions should take into account opportunities offered by distributed platforms, such as the Grid and P2P.

Recommendation 1.7

Encourage individuals and organizations to directly produce and make available born-digital material in the proper format and to make available digitized primary material that in the present system is not normally published, such as raw data and field notes.

- Much information is already born digital, reducing digitization costs into the future; proper formatting of such materials will make ingestion into repositories much easier.
- Currently, scientists and scholars are rewarded for their number of publications and their number of citations received as seen from a citation index. They evaluate primary content (data and cultural resources) and publish conclusions. This practice encourages hiding the primary material. Most citation indices even ignore electronic publications. Three measures can help:
  - Political pressure to change the evaluation criteria of scholarly and scientific work.
  - Establishment of public rewards for providing and curating high-quality digital material.
  - Instrumentation to trace the use of published digitized material as an additional citation index. Invest in the necessary monitoring systems, which can serve for accounting as well.

Question 2

What measures could be taken to promote private investments and new business models such as public-private partnerships for digitizing and making historical collections accessible?

Recommendation 2.1

A European Digital Library (EDL) should incorporate a service infrastructure that provides users with seamless access to fee-based information from publishers and other providers.

This service would manage payments to multiple fee-based providers in a pay-as-you-go fashion, including the possibility for micro-payments if the user needs access to small amounts of information. Such arrangement will benefit both users and fee-based providers who gain a new marketing tool. It would be entirely appropriate for EDL to collect a percentage from each transaction and thus establish an income stream that supplements public funding.

Technically, this can be accomplished in two ways (which can be implemented simultaneously):

- Provide access to publishers' sites through EDL's federated search, i.e., treat publishers' sites as collections covered by EDL. This works for documents that are available in digital form from publishers. Being able to market through EDL might encourage publishers to make available additional works in digital form, either by converting legacy digital files available from the production process or by applying OCR to non-digital material.
- Digitize copyrighted works in agreement with the corresponding publishers and provide fee-based access to these works.
- The same model could be used to encourage private third-party investment in digitization of in-copyright and out-of-copyright works. There are companies now that sell CDs of out-of-copyright works or provide fee-based Web access to them. An EDL could be a marketing and distribution tool for such companies in the same way as for publishers.

The following table is a first attempt at showing public and private involvement.

	Public role Government agencies, libraries, archives, museums	**Private role** Publishers, professional associations, telecommunication companies
Technical infrastructure **Network access** **Servers**	• Government agencies may create and maintain infrastructure • Government regulates private providers • Government may provide incentives for private providers	• Telephone companies, cable television companies, satellite access companies • Internet access providers
Digitization: Type of material	• Archival and museum materials • Out-of-copyright materials • In-copyright materials (by arrangement) • Born digital materials, made public	• Out-of-copyright materials (private companies digitize to sell digital copy) • In-copyright, digitized by owner • Born digital materials, held privately
Physical access **Content provision**	• Access to out-of-copyright materials • Portal to private access sites • Infrastructure for pay-as-you-go and micro-payments to facilitate access to fee-based materials • Subscription to private sites for a user community (For example, a university library may subscribe to a publisher's full-text journal site to give physical access to all its users)	• Access to in-copyright material . - Many publishers and professional associations provide fee-based full-text access to their journals and monographs - Movie distributors provide fee-based access to movies from many sources. - Same for music or image distributors) - Private fee-based portal sites ("digital book seller")
Preservation	• Digital deposit copies	• Private archives
Intellectual access, tools for use (annotation, processing), tools for collaboration	These services can be built on top of the content and physical access infrastructure as added-value services. They can be implemented by any combination of server-side software and client software.	
	• Public services • Intellectual access to all materials	• Fee-based services • Specialized user communities

Table 2.1. **A first look at public and private roles in a European Digital Library.**

Recommendation 2.2
Perform research in the value generation chain and identify parties that would have a natural interest to pay directly or indirectly for services.
Perform research in accounting models and pricing policies. Employ user-simulation techniques varying prices and paying parties. Disseminate results to organizations. Organize test beds for promising accounting mechanisms.

Question 3

What measures of a legislative, technical, organizational or other nature, could facilitate the digitization and subsequent accessibility of copyrighted material, while respecting the legitimate interests of authors?

The service infrastructure and business aspects of this question are addressed in Recommendation 2.1. The focus here is specifically on rights management.

Recommendation 3.1
Take legal and other necessary measures to establish easy-to-use clearing houses/collecting societies with authority to negotiate conditions for the use of materials.

- It is a huge task to clear all rights to material, especially when several people are involved, as in television programmes. Rights clearinghouses would ease this burden and, therefore, stimulate usage of copyrighted material. Payments should not necessarily follow the traditional models used for physical material, e.g., where payment is according to potential usage and the number of copies. Other means of identifying usage are available in the electronic world and they should be adopted.
- The Scandinavian model of extended collective licensing could be a suitable method for clearing rights to give access to digitized material. According to this model, it is possible, in particular cases specified by law, to make agreements with collecting societies that are binding also for authors and other right-holders, who are not members of the collecting society.
- The Creative Commons initiative gives an example where the licenses are available in machine-readable form. This should be the case for all licenses and rights in general. The EU should lead the way towards establishing a common European (and world-wide) initiative (could be in connection with MPEG21). As part of this initiative a legislative ontology should be developed.

Recommendation 3.2
Incorporate complete machine-readable rights metadata (policies, licenses, authorized user groups, etc.) into an EDL.

Question 4

Is the issue of orphan material economically important and relevant in practice? If yes, what technical, organizational and legal mechanisms could be used to facilitate wider use of this material?

Orphan material plays an important role and a method for handling the rights to this material is desirable.

Recommendation 4.

Establish a legal framework that allows for the legal use of materials when the corresponding rightsholder(s) cannot be identified. This can be accomplished by legislation at the European level or coordinated legislation in the member states.

- One possibility is to develop a model of 'preclusive claim', whereby after a diligent search to trace the rightsholder, interested parties might advertise their intention to digitize or otherwise use certain named works and be entitled to such use if no legitimate rightsholder comes forward and objects within a specified time (for example, 90 days).
- Another possibility is the use of an extended collective licensing mechanism.

Question 5

How could public domain material and other material available for general use (voluntary sharing) be made more transparent and widely known in order to facilitate its online availability for subsequent use?

This question has three aspects:

1. Improved searching (improved resource discovery)
2. Improved user education and marketing
3. Improved rewards for the curators of this material for providing access (which has already been captured in Recommendation 1.7)

Recommendation 5.1

Based on existing work, select, modify, or produce a full-functioned search/access engine, specifically geared towards cultural-heritage materials. The architecture of this search engine should allow for the easy addition of special functions for other domains, such as scientific or business materials resulting in one underlying search engine with multiple interfaces, each adapted to a domain or user group.

A European Digital Library will provide access to born-digital or digitized content in many formats (text, images, sound, multimedia, interactive objects, etc.) held by public institutions such as libraries, archives, and museums, as well as private publishers, such as professional associations and telecommunication companies. Searching and retrieving such material using a general-purpose conventional search engine, however useful in many cases, does not suffice in order to help unlocking the full value of cultural-heritage or scientific content.

Multimedia search/querying and indexing engines are still not mature enough to be used in digital libraries. Queries on metadata of digital objects are routine, e.g., find photos whose title contains the keyword "Titanic", but for many multimedia objects metadata are not available. For digitized multimedia objects one can search on the contents of digital objects, e.g., find photos looking similar to this photo (of the Titanic). While state-of-the-art techniques (image/pattern matching) would have some success with content-based searching, the results would be below expectations and query performance would probably be poor (there are numerous mathematical calculations involved in answering such queries). Hence, for long-run evolution of digital object repositories, the following are needed:

- creation of digital repositories of detailed distributed metadata information for multimedia objects
- methods for automated or computer-assisted creation of metadata for multimedia objects.

- query languages to express queries on contents and metadata of digital objects (audio / video / object shapes, etc.)
- indexing structures and query processing and optimization algorithms to speed up query execution
- personalization techniques for searching to adapt to the searcher and to the changing mechanisms of access (e.g., PDA, etc.)
- appropriate user interfaces for digital content queries, including example-based queries
- querying mechanisms that allow for unifying metadata and content summarization under a common semantic model in order to retrieve information related to a specified context of interest.

An EDL should provide users with capabilities for performing intelligent searches enabling them to better understand, appreciate, and re-use stored objects. It should manage highly effective semantic retrieval, taking into account domain knowledge and providing new ways to intelligently search and retrieve objects based on such knowledge.

Recommendation 5.2

Looking into the future, take steps towards the integration of access and retrieval with scientific and business workflows, so that digital content is readily reused and relevant results are added to the scientific and scholarly memory.

- Suitable formats and standards for interoperability of inputs, outputs, and metadata need to be developed and adopted. Traceability of content provenance is key in this context, as is the possibility of recalculating results with updated parameters or enhanced input.
- Interdisciplinary working groups and studies are needed that generalize over domain-specific practices of scientific workflow and identify appropriate query paradigms to respond to retrieval requirements as they appear in characteristic stages of work and processing.

Recommendation 5.3

Once a European Digital Library is under development, conduct user education at all levels and an informative marketing campaign to make users aware of the richness of its content and the ease and sophistication of its access mechanisms.

- Training is necessary at the elementary and secondary education levels to prepare a knowledgeable user base that can fully exploit the opportunities offered by the system.
- For this purpose, a careful assessment of competencies followed by curriculum development is needed.
- An EDL should be proactive in making users aware of its resources, e.g., by having a spotlight on the digital collection of the month.

Preservation of digital content

Question 6

What priority measures – in particular of an organizational and legal nature – should be taken at national and European level to optimize the preservation of digital content with the limited resources available?

Digital preservation normally entails all activities that ensure collection of, maintenance of, and physical access to documents or other resources over time, and proper rendering and interpretation of resources once accessed. Digital curation, an essential aspect for preservation, supports integrity, authenticity, reliability, security, maintenance and access to digital materials across time and systems. Preservation and curation include collection policies and awareness raising on the ingest side. Ingest and access are today governed by national laws, and solutions vary from country to country and from institution to institution. From a technical side, however, the problem of digital preservation is international and so should be the solution.

Recommendation 6.1

Define standards and frameworks to stimulate the implementation of sound preservation strategies.

This has both technical and organizational consequences and, therefore, the recommended activities fall in these two groups. Some of the recommendations are direct quotations from "Invest to save"[3] and are put in "..".

- Coordinated awareness-raising.
- "Building up expertise: There is a need to build up expertise concerning preservation of digital objects, especially in small, highly-specialized companies. At the moment ad-hoc solutions for each specific environment predominate."
- Framework for service infrastructure: The definition and implementation of a framework, which will allow different activities to take place, could stimulate public and private sectors to develop applications for tailored services, e.g., format migration or rendering.
- "Developing an organizational framework: Based on the introduced organizational methods for preserving physical objects in filing departments, libraries, and archives, new methods reflecting the special requirements of born or converted digital objects need to be developed."
- Establish best practices to show how preservation can be included as an integrated part of the life cycle of objects. This is especially relevant for "large, distributed organizations such as the broadcast industry", as they "have no common rules on how to describe the archiving of digital objects. As a result, archives within organizations exist where no information exchange and re-use can be realized. Overcoming this deficiency requires technologies that enable integration of existing digital objects."
- "Developing cost modeling tools: Ensuring organizational support for preservation and enabling longer-term planning for the revenue implications of engaging in

[3] Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation 2003

preservation depends upon the availability of cost modeling tools. These are currently lacking."

Recommendation 6.2

Establish a clear legal framework that mandates ownership of digital objects to the producer.

As said in "Invest to Save", "Legal issues could become one of the major obstacles to introducing long-term preservation. Only in the presence of clearly defined rules and policies, which mandate ownership of digital objects to the producer, will producers be willing to participate in the preservation process of their digital objects."

Recommendation 6.3

Ensure the capture, storage, and use of provenance and related metadata

An important issue for both digitized and born-digital content is to consider data provenance as high-class information. Usually, a series of production steps is necessary so that a document can be fully integrated into a digital library (e.g., digitization – if not born digital–, extraction of features, continuous enrichment with annotations, linking it with other content, etc.). Data provenance not only allows one to exactly trace the process of digitization, but also reflects the full lifetime of an object. This is even more important for certain born-digital content such as, for instance, parts of scientific digital libraries. Consider data derived from experiments in physics or in life science. Here, in addition to the workflow used for producing the object, it is also of high importance to include in the provenance information all details on the configuration of the machinery used in order to be able to reproduce the experiments. For compound objects, data provenance needs to be considered for the constituent parts and the composite object independently. Therefore, the following actions are necessary:

- Raise awareness for complete and thorough production of provenance information during the digitization process but also during the complete lifetime of a digital object.
- Establish a list of mandatory metadata to be provided during production, digitization, enrichment, and combination of digital content.
- Include in the provenance metadata details on copyrights as well as on the responsible organizations/individuals for each single step in the production process of a digital object.
- For digitizing organizations: Derive (parts of) metadata automatically during the digitization/production process (including digitization and OCR method and parameters, software used, and standard provenance metadata).

Recommendation 6.4

Ensure that a European Digital Library contributes to and conforms with international standards for unique identifiers and linking structures

Information is accessed on a global scale, and "national boundaries" become blurred as a result. Also, as technologies develop, the division between what has been published and what has not been published becomes more and more theoretical. This has implications for strategies, both related to the collection of material and for strategies for future access. As an example, consider material on the Web, where for the user, it is unimportant in which country a given server is placed or in which country a given document is posted. For future access to

the stored objects in any digital repository or the Web at large, it will be impractical and not understandable if the link infrastructure does not work. The following two activities follow directly from the above:

- Support for persistent unique identifiers: A growing problem is the lack of unique identifiers on a national and international level. Several schemes are available but few outside the commercial publishing world are using them. Recommendations for unique identifiers (not necessarily limited to one scheme) and resolution services to support the unique identification of present as well as past material are needed at the European level.
- Cross national link infrastructure: A challenge is to retain the international link infrastructure, which is such an important part of the Internet. It will not be understandable for users in the future if they can only access the material from one country. It is therefore important to establish an infrastructure that allows links to be followed across natural and temporal boundaries. The unique identifiers mentioned already will be one of several ingredients in such an initiative.

Recommendation 6.5

Install service centers for monitoring technology obsolescence to inform organizations about obsolescence risks, to provide migration services from obsolete data carriers or software, and to preserve obsolete technology as long as possible or develop emulation methods.

Recommendation 6.6

Foster formation of community-specific consortia for the maintenance of copies of digital objects at multiple, distant sites as a measure against physical loss. These systems should maintain enough valid copies at multiple sites via network access, following security specifications and monitoring risk of loss per object.

- For instance, art museums could collaborate to hold each others' copies of their huge high-resolution multispectral images of art works.
- Each domain has different requirements. For instance, a natural history collection would digitize small, distinct features of specimen, such as hairs of plants, in high resolution.

Question 7

Is there a risk that national legal deposit schemes lead to a multiplication of requirements on internationally active companies? Would European legislation help avoiding this?

According to the Universal Availability of Publications (UAP) principle, each country is responsible for its own publications. This may result in additional burden on companies if the material in question is deemed published in several countries. With the increasing cooperation among European National Libraries, it is most important that a copy of every publication be deposited in one European country; and the question in which country the copy is deposited becomes less significant.

Recommendation 7

Initiatives should be taken at the European level to ensure that a concept of "main association" is introduced in the law. The idea is that multinational companies have a way to determine for any given application a single deposit point. In other words, when by present legislation a company has to deposit copies of a publication in several European countries, by new legislation it could determine the country that is predominant and deposit a copy there.

Question 8

How could research contribute to progress on the preservation front? Which axes of work should be addressed in priority by the forthcoming Specific Research Programmes as part of the 7th Framework Programme?

The solution to a number of the areas identified above as organizational challenges requires research to provide satisfactory answers. Research areas identified in "Invest to Save" are still valid: "The digital environment is fundamentally reshaping how society is producing, disseminating, using, and repurposing information and knowledge. This transformation requires effective digital archiving solutions as part of the infrastructure for a knowledge-driven economy. The attempt to replicate traditional mechanisms for appraising, acquiring, documenting, and managing information in the digital environment has not provided mechanisms that respond to the complexities and fluidity of digital entities themselves and their contexts. While acknowledging the value of many conventional archival principles, the Working Group concluded that archival processes must be redesigned and re-engineered. This change will require a paradigm shift in research if it is to provide the innovation, whether theoretical, methodological or technical, necessary to enable long-term access to digital materials."

Based on the analysis in the report as well as the issues raised in the above comments, one can identify five research areas:

- Socio-economic studies of the processes that are being preserved and of the methodology of preservation adopted by archives. The studies address areas such as cost models, social behavior, selection/appraisal, life cycle management, and impact of electronic communication (such as new ways of mass communication, research communication and e-government).
- Automatic ingest processes, including automatic metadata extraction, validation and authenticity checking.
- Automating the maintenance process and identify models for how this can be done
- Models for access and for international linking between relevant material residing in different archives
- Highly parametric models for risk of physical loss of data carriers due to media failure, handling errors, disasters, etc. can be used for decision taking on the most cost-effective selection of storage media, storage spaces, and distribution strategies

2 Short-term actions

Recommendation 9

Bridge the gap between research and application through preparing practice-focused, solid, and exhaustive state-of-the-art reports and tool repositories and through establishing competence centers that are accessible through a one-stop portal. This is a very short-term need and should be funded now, if at all possible, to provide a better basis for high-quality efficient digitization and high-quality access

Now is the moment to transfer the existing research results to the applications. This should be **funded now** without waiting for FP7 so that the results are available before large-scale digitization and other processing starts under the new program.

- Many research results and system prototypes exist. Apply them to building a European Digital Library as planned by the Commission.
- Support focused research to fill in gaps faced by an EDL initiative now.
- Do not neglect research that deals with issues in the next step so that the solutions are ready before the next step is taken. For example, foster studies of domain-specific access and interfacing requirements and conduct experiments for advanced scenarios of integrated DL access and scientific workflow management.
- Use the systems being created now as a test bed for new methods.
- Encourage more interdisciplinary collaboration in European funded research. Evaluate research proposals by truly interdisciplinary teams.

The state-of-the-art reports should be structured so that they can be read in context and used as a searchable knowledge base with links into the literature for background and further detail. They should be searchable from the perspective of practice, supporting policy decision as well as the nitty-gritty detail of digitization, preservation, and access. They should be structured in such a way that they can be combined into one comprehensive knowledge base.

Preparation of the state-of-the-art reports should take advantage of any published or unpublished surveys already existing. They should represent one-stop source of advice to practitioners.

The tool repositories should be tightly linked to the state-of-the-art reports so that guidelines are accompanied by links to tools that can aid in their implementation.

State-of-the-art reports are needed in the following areas. Overlaps between these reports need to be identified so the teams preparing these reports can coordinate accordingly.

9.1 Tools for digitization, OCR, and other related technologies.

- Evaluation of existing digitization work (by major stakeholders, national programmes, etc.) with respect to technical, legal, usage, and other aspects. Some of the stakeholders have been carrying out such evaluations already (e.g., German national efforts).
- Survey of digitization techniques, especially for multimedia objects, three-dimensional objects, composite objects, and fragile objects.
- Survey of best techniques in OCR in different fonts and different languages and of different types of materials, such as musical scores (see, for example, GAMERA, http://ldp.library.jhu.edu/projects/gamera/).

9.2 Preservation issues, strategies, techniques

See the recommendations under Question 6 for the issues involved.

9.3 Access, search engines, annotation, collaboration, data integration

Particular emphasis should be put on access to cultural heritage materials in diverse formats

9.4 Metadata standards for all kinds of objects

See Recommendation 1.4.

9.5 Multilinguality

Most of the basic building blocks for multilingual information processing exist already; in particular, the "enabling technology" is there and known, e.g., encoding and representation issues (Unicode), language identification, localization tools, and tools for basic indexing (stop words, stemmers, morphological analyzers, named entity recognition, etc.).

It is difficult, however, for the application communities to have a clear picture of what they need / what is available and how to use it, and the tools need to be validated so they can be recommended to non-specialists. There needs to be an inventory of these tools with indication of what tools are suitable when and also an indication of what is missing. A "one-stop shopping" repository will be very useful, where people can easily obtain the tools that are appropriate for their requirements, including proprietary tools that need to be bought (possible through a license acquired by the European Digital Library).

Thus, the report on multilinguality should include the following:

- guidelines and how-to-do-it advice for dealing with any language
- a catalog of tools indexed by purpose and language, with pointers to where to get them
- a guide to methods for speedy development of tools such as stemmers for languages for which a given tool does not exist.

The report should consider all languages important in Europe, including historical languages.

9.6 Multimedia and other complex objects

See Recommendation 1.6.

Recommendation 10

Develop test beds for the evaluation of DL Research Results

Results of research and development of DL functionality and services should be incorporated as soon as possible into appropriate test beds, thereby gradually improving and enriching the latter. This allows for early evaluation of DL functionality. In addition, it supports the integration of new functionality with existing services. These activities should not be restricted to a single test bed but rather consider different test beds by which different aspects can be evaluated. In the short term, TEL should be one of the test beds considered. Having TEL as a large-scale test bed will not only make new functionality (such as faster search, query expansion, cross-language search, full-text search, integrated result presentation personalization) available to the TEL user community, but will also support the evaluation of new functionality based on the TEL content. While the TEL test bed activities are content and functionality-oriented, other test beds should focus on infrastructure aspects, since the EDL will have large volumes of objects and a high number of concurrent users running complex DL services or workflows. Therefore, an underlying infrastructure that is highly scalable and that efficiently uses available resources (e.g., by relying on grid infrastructures and/or peer-to-peer concepts) is needed. BRICKS and DILIGENT are examples for such infrastructures.

3 General principles for the development of a European Digital Library

The overall goal of the initiative to create a European Digital Library (EDL) is to achieve a more inclusive European Information Society and to overcome the geographical and social digital divide. Content must be brought to the citizens in an easy, efficient, and affordable way.

Recommendation 11.

Maintain a short-term focus while keeping in mind a long-term vision

It is a highly useful and necessary first step to focus on large-scale digitization and physical access to materials. This is doable now and lays the foundation for providing more advanced services later, as discussed in Section 4 of this report. **It is important to keep these advanced services in mind and make sure at every step that results produced now are compatible with future functions**. For example, document formats must allow for adding annotations so that a collaborative annotation system can be built on top of a static digital library.

Such a strategy ensures near-term success in mass digitization and collection building, yet enables the gradual development of a powerful access and use system that uses computer and data communication power for full vertical integration of user functions (retrieving information in several media and formats, learning, editing, annotating, sophisticated processing, collaborating, taking account of the user's present context and of sensor-generated data) and horizontal integration across national, linguistic, and cultural boundaries. In the long term, this will be the hallmark of the European effort, setting it apart from simplistic search engines. The latter are useful for many purposes but do not meet the increasing demand for more sophisticated functionality, with intelligence behind the scenes to better support the user in focused searches and use of the materials found.

Ultimately the system being developed must support use and exploitation of information as an integral part of application workflows. This goes beyond access. This is the area that, in the long term, presents the strongest challenges and that needs the development of fresh ideas on how to address it adequately. Use and exploitation should dictate the design and implementation of all system components. This requires detailed studies of users. One could view a European Digital Library as a system that ultimately provides specific services to specific user groups rather than from the distant perspective of a general-purpose library and its traditional overall approach.

In concerted actions, key scientific and business areas should be selected for pilot implementations of integrated services, with the purpose of identifying more and more generic new services. Interdisciplinary design and evaluation teams should ensure the tight coupling between DL R&D and application needs.

This strategy of short-term practical results and long-term incremental development of a system that fulfills the promise of the Information Society will put Europe in a strong competitive position.

Recommendation 12.

Apply a user- and task-centered design and development strategy

- Design and development should be based on a thorough understanding of the user communities and their tasks, gained through user studies and analyses of present and future requirements. From these, scenarios and user personas can be developed, which in turn will lead to prototypes. Prototypes should then be tested with users through usability and effectiveness analysis. Throughout the process, users should be helped to imagine the possibilities of the new technology so that the vision of new systems is not hampered by the limited false impressions of users on what can be done.
- User studies are essential for the mid- and long-term development of a fully-functional European Digital Library but they need not hold up the short-term digitization of collections (as long as the digitization and data formats take into account multiple future uses). User studies are important for designing the services that build on the foundations of digital collections.

Recommendation 13.

Unify access to libraries, archives, museums, and other memory institutions

A European Digital Library must provide **unified access to materials from libraries, archives, museums**, other memory institutions, and ultimately, the Web. The end user does not care about the difference.

Recommendation 14.

Establish a unified service-oriented framework for all digital library projects sponsored by the European Commission as part of one European Digital Library system.

There should be **a unified framework for considering the issues arising in all digital library projects sponsored by the European Commission**, in particular, for the cultural heritage digital libraries and the scientific digital libraries. This applies to issues of technical infrastructure as well as of services that build on the infrastructure. Special requirements can be handled by plug-in modules.

A major requirement for a future EDL is that it is based on an infrastructure that allows seamless integration and combination of different DL services. By this, DL functionality can independently be designed and implemented by means of specialized services. The service-orientation paradigm then allows for the application-specific combination of services as building blocks into processes or workflows. An EDL has to support mechanisms for easy combination of services and for the verifications of newly designed compound services. Finally, also the reliable execution of compound services (including failure handling) has to be supported by the EDL infrastructure.

In addition to the underlying digital-library infrastructure, the core building blocks and services have to be identified. To this end, a *reference model for a Digital Library Management System* is needed that lists mandatory and optional DL components and that describes the functionality of these components within a conceptual framework. This has to consider both end-user functionality and system functionality that, transparently to the end users, provides added-value both to objects and application-oriented services.

Recommendation 15.
Aim for a federated system of interoperable partners with one-stop access portals for specific user groups and uses

- Move from digital libraries as integrated, centrally controlled systems to dynamic, configurable federation of DL services and information collections**.**
- We see a European Digital Library as a comprehensive system subsuming TEL in both content and functionality, with TEL as a component and as a large-scale test bed for developing and testing advanced functions.

Recommendation 16.
Ensure interoperability on all levels: technical infrastructure and data formats, descriptive metadata, and Knowledge Organization Systems (KOS)

Recommendation 17.
Provide mid-term and long-term funding of research and development that is focused on making the long-term vision a reality.

4 Vision for the future of EDL

The DELOS vision is that Digital Libraries will become the universal knowledge repositories and communication conduits for the future, common vehicles by which everyone will access, analyze, evaluate, enhance, and exchange all forms of information. They will be indispensable tools in the daily personal and professional lives of people. They will be accessible at any time and from anywhere, and will offer a friendly, multi-modal, efficient, and effective interaction and exploration environment. Efforts towards this vision require significant changes in the present Digital Library development strategies, with respect to functionality, operational environment, and other aspects. There is the need to overcome the major limitations observed in the development of present-day systems, which are essentially "content-centric" and "one-of-a-kind", i.e. each system has been developed having in mind a specific content, a specific user community, and a specific application.

Implementation of the vision requires instead that Digital Library systems have the following characteristics:

- *Person-centric efforts*: Humans are at the center of Digital Libraries and all efforts to develop them should be initiated and motivated by needs to provide interesting and/or novel experiences to users. Furthermore, Digital Library systems should synthesize all information that is available about each person in a cohesive whole, so that they may offer personalized treatment to individuals or classes of individuals based on their profiles.
- *Communication-centric & collaboration-centric functionality*: The main role of Digital Libraries must be to facilitate interaction of scientists, researchers, professionals, government and business workers, and the general public on themes that are pertinent to the information stored. Storage of this information and access to it is only a small (although still essential) part of such functionality.
- *Generic technology systems*: For economy of scale, reusability, and extensibility, generic Digital Library Management Systems (DLMSs) should be developed that capture all common management aspects of Digital Libraries. Supporting any further, environment-specific needs on content manipulation or user interfaces should be developed in a customized fashion on top of DLMSs.
- *Maximum-reuse efforts*: Given the existence of industrial-strength DLMSs, every development effort should depend on them, avoiding much mundane work that is currently necessary, and should only focus on the specialized parts.
- *Globally distributed systems*: Digital Libraries should be managed by widely distributed systems, through which information sources across the world get interconnected to exchange and integrate their contents.
- *Universality of information and application*: Digital Libraries should be put in the service of ``all'' applications and should comprehensively manage ``all'' forms of content, from data to information to knowledge.

Some aspects of this grand vision are discussed in the remainder of this section. The structure is different from the previous sections, with more emphasis on general issues and less emphasis on specific recommendations.

4.1 EDL as a comprehensive system that supports users' work and daily activities

A future EDL should provide functionality to

- go beyond serving research and scholarship and education and **also support practice** (in medicine, in law, in business, in government)
- **support new ways of intellectual work:**
    1. information access should be embedded seamlessly into an integrated system that supports all of a user's work, information access as well as information use, information application, and derivation of new thoughts
    2. systems should go beyond paper-based limitations and associated metaphors that trap the user into old thinking
- **support collaboration, especially collaboration within communities of practice**

4.2 Ontologies as a core component of DLs, collaboration, and large-scale content integration

Ontologies are essential for efficient metadata integration, for developing and applying document tagging schemes, for structuring documents, for subject access, and for reasoning with data. Beyond that, a well-constructed core ontology (or common semantic model) of relationships would provide a framework for incrementally building large distributed knowledge bases through ontology-supported linking and integration of many disparate assertions, many of which can be extracted from documents. Hence, many users can use ontology-driven tools for computer-assisted extraction of data from text for their own purposes, making their work more efficient. As importantly, these data can then be reused and made useful for a large community: It can be incorporated (connected, interlinked) into a large distributed knowledge base to be used for sophisticated retrieval, reliable question-answering, and reasoning across domains. This would provide a big boost to Semantic Web technologies.

Building ontologies is expensive, requiring much intellectual effort. In some cases, extraction of ontology information and/or automatic classification can help. The core ontology mentioned above needs to be developed centrally by a group of experts; there may have to be several computing (but still to some extent relatable) core ontologies. For other purposes, a single ontology developed collaboratively by a community of practice may work. Moving in the direction of such "Community-Developed Ontologies" is based on the following premises:

- Ontologies are difficult to establish
- Top-Down "authority"-based ontologies in most cases do not work
- Communities (of expert users) can build their own ontologies
- Different communities can build different ontologies for the same library
- Methods, techniques and tools are needed

Ontology and terminology services can support system designers, indexers / catalogers / metadata editors, and end users.

4.3 Personalization

As ever-increasing amounts of information become available to a growing number of users, it becomes very difficult for users to find information they need. Moreover, different people with different background, goals, interests and preferences may expect a different,

personalized system behavior. What distinguishes a personalized system from a traditional information access system is essentially the existence of user profiles that store information based on models of users either as individuals or as members of groups. Stored user information is used by the system in order to adapt its behavior to the needs, characteristics and preferences of users.

These profiles might be specified explicitly or be automatically derived from the history of interaction of a user with the DL. Individual profiles can also be used for continuous queries, i.e., queries that are defined once and are continuously re-evaluated. In general, system behavior may be adapted, i.e., personalized, at different levels: at the content selection, content presentation, services, or interaction level, taking into account the goals, interests, and other characteristics of the users. For example, different users are provided with different content according to their interests and preferences. The same content can be presented to different users, in a summarized, or an extended form or in different layout and colors depending on the user. Different users may have access to different services, which may be customized according to the needs and preferences of the users targeted.

A future EDL should provide functionality to

- enable people with varied goals and characteristics to access an ever-growing amount of born-digital or digitized information with the minimum cognitive load
- explicitly and implicitly derive user preferences and profiles
- personalize information access to digital content at various levels, e.g. content selection, interaction level and so forth, and in a transparent way, taking into account user profiles

4.4 Pervasive information, mobility of users, location and context awareness

Digital information is increasingly becoming ubiquitous and pervading more and more everyday life. With the proliferation of mobile devices, access to this information is available on a 24/7 basis. However, mobile devices and mobile information access also raises new challenges.

First, the user interfaces of mobile devices are limited. This means that content might have to be delivered in different quality in order to be displayed on a mobile device (e.g., in lower resolution, compressed, or summarized). The same is true for the format in which content is made available to mobile users.

Second, the fact that the power supply is limited and mobile devices cannot be continuously online raises demand for other types of queries (asynchronous interaction, publish/subscribe-based information access). Interaction patterns between a mobile device that frequently disconnects and reconnects and a DLMS are influenced. This is even more important when the mobile device holds content that is being made available via an EDL.

Third, other major challenges are imposed by his/her demand for context and location awareness. Depending on the location of mobile users, their individual information needs may change. This does not only affect the content to be accessed but also the functionality to be used (e.g., consider a user being located at a historic site; a particular demand could be information, back in time, on the events that have taken place at this particular location). Similarly, also the context of digital-library users might influence their particular information needs and the way digital-library content is accessed.

A future EDL therefore has to provide support for the detection of user location and context and has to consider this information for information access from mobile devices.

A future EDL should provide core functionality to

- access content in different quality
- deliver content in different formats
- make the results usable on different (mobile) devices
- consider context and location

4.5 Data streams

Data streams are continuous streams of information, stemming from either software or hardware sensors, to be integrated into a DL. Streams are particularly important in scientific digital libraries (data from sensor networks as they are used, for instance, in earth observation but also in diverse other applications). However, streams are also of primary importance for tracking mobile users and for detecting their current location and context (while, at the same time, respecting the privacy of users, i.e., the context and location is tracked only for the users that explicitly agreed to that). Due to the fact that data streams are in general objects of infinite length, it is important to provide appropriate mechanisms for online processing of (windows of) data streams. In addition, aggregation of stream information and storage of aggregated stream data for later exploitation (especially in scientific applications) are needed.

A future EDL should provide functionality to

- process continuous streams of data (e.g., allowing for customized stream applications by combination of different stream operators)
- integrate data streams and aggregated stream data into the DL, and provide access to this data

Appendix 1. List of Participants

On December 5 - 6, 2005, the DELOS Network of Excellence on Digital Libraries held a brainstorming meeting in Nice, France, to formulate responses to the i2010 Digital Library questions from the European Commission and to discuss the DELOS vision of the future of Digital Libraries. Participants came from within and without DELOS, from Europe and the United States, including librarians, researchers in the DL field, and representatives of the European Commission.

Last Name	First Name	Affiliation
Agosti	Maristella	University of Padova
Bolchini	Davide	University of Lugano
Borbinha	Jose	INESC-ID
Bury	Stephen	British Library
Casarosa	Vittore	ISTI-CNR
Castelli	Donatella	ISTI-CNR
Chambers	Sally	The European Library, Koninklijke Bibliotheek
Chailloux	Jérôme	ERCIM
Christensen-Dalsgaard	Birte	State and University Library
Christodoulakis	Stavros	Technical University of Crete
Clavel-Merrin	Genevieve	Swiss National Library
Cousins	Jill	IKJE bibliotheek
de Vries	Repke	Koninklijke Bibliotheek
Doerr	Martin	FORTH
Dujacquier	Isabelle	Ministère de la Communauté française de Belgique
Forster	Horst	European Commission
Freyre	Elisabeth	Bibliotheque Nationale de France
Fuhr	Norbert	University of Duisburg-Essen
Griffin	Stephen	National Science Foundation
Ioannidis	Yannis	University of Athens
Koch	Traugott	UKOLN
Le Dantec	Bruno	ERCIM
Manson	Patricia	European Commission
Meghini	Carlo	ISTI-CNR
Neuhold	Erich	University of Vienna and Research Studios Austria -
Paolini	Paolo	Politecnico Milano
Peters	Carol	ISTI-CNR
Picininno	Marzia	Ministero Beni Culturali
Ross	Seamus	University of Glasgow
Schek	Hans-Joerg	UMIT
Schuldt	Heiko	UMIT
Sellis	Timos	School of Electrical and Computer Engineering
Soergel	Dagobert	University of Maryland
Solvberg	Ingeborg	NTNU
Thanos	Costantino	ISTI-CNR
Tudhope	Douglas	University of Glamorgan
Weikum	Gerhard	Max-Planck Institute for Informatics
Zumer	Maja	University of Ljubljana

Appendix 2.

Ad-hoc faceted classification of digital library issues

This is a proposal for an ad-hoc faceted classification for the organization of issues arising in the European Digital Library initiative that emerged in the DELOS brainstorming process. It should be useful for others who think about this problem and plan for the EDL. The facets are strictly pragmatic based on the way the concepts combine. The classification makes no claim for completeness or completely thought-out conceptual structure. It was drafted very quickly simply for the purpose of organizing a set of ideas.

Outline

A **Facet A. General themes across functions**

A1 . Total systems approach
A2 . Evolution of DLs: Integrated information and task environments
A3 . Management. Economic and legal questions. User studies. Evaluation
A4 . Coordination, collaboration, interoperability
A5 . Research
A6 . Education and public relations

B **Facet B. Architecture, ontologies, metadata, document processing**

B1 . Architecture/framework/infrastructure
B2 . Document models, document structure
B3 . Ontologies. Knowledge Organization Systems (KOS). Other authority systems. Metadata schemes.
B4 . Metadata
B5 . Document processing
B6 . Linking and content integration

C **Facet C. Function: Digitization, preservation, access**

C1 . Collection development, selection, appraisal
C2 . Digitization / OCR / digital formatting
C3 . Preservation and long-term access. curation
C4 . Access and use

Facet D. Type of material

D1 . Types of collections / systems by number of languages
D2 . Types of collections / systems by degree of control
D3 . Material by medium
D4 . Material by complexity
D5 . Material by interactivity
D6 . Material by static vs dynamic
D7 . Material by publication status
D8 . Material by copyright status
D9 . material by subject area

A **Facet A. General themes across functions**

A1 **Total systems approach** - having the ultimate goals in access and use govern the solutions for digitization and preservation

A2 **Evolution of DLs: Integrated information and task environments**

A3 **Management. Economic and legal questions. User studies. Evaluation**

A3.1 . **Political issues surrounding digital libraries**

A3.2 . **Economics of DL**

A3.2.1 . . Economic and financial issues

A3.2.2 . . Business models including public and private partners
The commission launches i2010 in order to boost the digital economy. Better business-models.

A3.2.2.1 . . . Reward models and business models for large-scale collaboration on content integration.
New roles. BT A4.2 Collaboration

A3.2.2.2 . . . pay per view

A3.3 . **Legal questions** and their economic ramifications (ref. Google projects)

A3.3.1 . . Intellectual property rights, copyright, digital rights

A3.3.1.1 . . . Digital rights management

A3.4 . **Authenticity, integrity, trust, security, and privacy**

A3.4.1 . . Authenticity

A3.4.2 . . Integrity

A3.4.3 . . Reliability

A3.4.4 . . Trust

A3.4.5 . . Security

A3.4.6 . . Privacy

A3.5 . **Organizational questions**. Cooperation between European organizations
Digitization on a grand scale will in most countries be the responsibility of the National Libraries and other national Archives and Agencies, - in cooperation with research, industry and businesses.

A3.6 . **Ensuring technical excellence**

A3.6.1 . . Narrow the gap between research and application. Disseminate newest technologies

A3.6.1.1 . . . Foster application-oriented research

A3.6.2 . .. Provide technical expertise through competence centers, preferably accessed through one central clearinghouse(through an excellent Web site as well as in-person consultation) BT A4.1

A3.6.3 . . Pilot projects to demonstrate new technologies

A3.7 . **Users and user studies**

A3.7.1 . . User groups

A3.7.2 . . User studies methodology

A3.7.2.1 . . . User studies / Ethnographic /

A3.7.2.2 . . . Focus groups

A3.7.3 . . Actual or planned user surveys, data from user surveys

A3.8 . **Evaluation. Quality criteria**
A3.8.1 . . Success criteria & measures
A3.8.2 . . Evaluation & testing methods
A3.8.3 . . Test-beds
A3.8.3.1 . . . Large multi-media repositories as test beds
A3.8.3.2 . . . Libraries/museums as technology test beds
A3.8.4 . . Prototypes RT A7

A3.9 . **Quality control, quality enhancement**
A3.9.1 . . Audit and certification
A3.9.2 . . Data cleaning

A3.a . **Costs**

A4 **Coordination, collaboration, interoperability**

A4.1 . **Coordination** - a key ingredient for success NT A3.6.2
A4.1.1 . . Setting and enforcing standards to ensure interoperability
Standards in specific areas of content, use, and users
A4.1.2 . . Sharing tools

A4.2 . **Collaboration**
A4.2.1 . . collaboration among providers
A4.2.2 . . collaboration among users

A4.3 . **Interoperability and integration**
A4.3.1 . . Interoperability RT A4.4
A4.3.1.1 . . . Mappings, transformations, crosswalks
A4.3.2 . . Integration of heterogeneous systems (DLs, ontologies, etc.)
A4.3.3 . . Ontology-driven interoperability and integration BT B3.3

A4.4 . **Standards** RT A4.3.1

A5 **Research**
A5.1 . Short-term research agenda (acquire knowledge and develop or improve techniques that are needed now)
A5.2 . Long-term research agenda

A6 **Education and public relations**

B ` **Facet B. Architecture, ontologies, metadata, document and knowledge p**

B1 .**Architecture/framework/infrastructure**

B1.1 . Overall DL configuration

B1.1.1 . . DL as an integrated, centrally controlled system

B1.1.2 . . dynamic, configurable federation of DL services and information collections.

B1.2 . Service infrastructure model

B1.2.1 . . Framework to support services in connection with library services

B1.3 . DL reference model

B1.4 . Repository design, repository model

B1.5 . DL software

B1.6 . Automation of processes

B1.7 . Scalability

B2 **Document models, document structure**

B2.1 . Document models

B2.1.1 . . Dynamic documents

B2.1.2 . . Composite documents

B2.1.3 . . Self-describing & self-monitoring entities

B2.2 . Document life cycle

B2 . Document formats

B3 . Document tagging schemes

B2.4 . linking information

B2.4.1 . . Beyond typed hypertext links: Advanced (discourse specific) linking models: linki reference (such as friend-of-a-friend, FOAF), by factual relationships, by cross-c relationships. Use of knowledge extraction to populate linking models.

B3	**Ontologies. Knowledge Organization Systems (KOS). Other authority systems. Metad**
B3.1	. Types of KOS by content NT B4.1.1
B3.1.1	. . Core ontologies, foundational ontologies, upper level ontologies
B3.1.2	. . Subject KOS
B3.1.2.1	. . . Concept systems
B3.1.2.2	. . . Terminology systems
B3.1.3	. . KOS for places and place names (gazetteers)
B3.1.4	. . KOS of events and historical periods
B3.1.5	. . Representation of temporal information, passage of time
B3.1.6	. . Authority list of person and organization names
B3.1.7	. . KOS of relationship types
B3.2	. Vocabularies of different levels of formality
B3.3	. Uses of KOS NT A4.3.3
B3.3.2	. . Connecting KOS into distributed global networks of curated knowledge; processin referential integrity
B3.4	. KOS registries NT B4.1.2
B3.4.1	. . Registries of subject and related KOS
B3.5	. Tracking KOS over time

B4 **Metadata** NT B3.1.7 Metadata schemes B3.4.2 Metadata registries
B4.1 . Metadata schemes, metadata registries
B4.1.1 . . Metadata schemes BT B3.1
B4.1.2 . . Metadata registries BT B3.4
B4.2 . Types and uses of metadata
B4.2.1 . . Provenance
B4.2.2 . . Descriptive metadata
B4.2.3 . . Subject metadata
B4.3 . Level of metadata
B4.3.1 . . Document-level metadata
B4.3.2 . . Collection description
B4.4 . Processes with metadata
B4.4.1 . . Automatic classification, metadata extraction, feature extraction BT B5
B4.4.1.1 . . . Metadata extraction
B4.4.1.2 . . . Automatic classification
B4.4.1.2.1 . . . . Feature extraction
B4.4.1.2.2 . . . . Automatic classification algorithms
B4.4.1.3 . . . Automatic assignment of geographic coordinates
B4.4.1.4 . . . Automatic appraisal BT B4.3.1, C1.2
B4.4.2 . . Metadata capture, manual metadata creation
B4.4.2.1 . . . Metadata capture during document creation and from user actions
B4.4.2.2 . . . Manual metadata creation
B4.4.2.3 . . . Metadata creation tools
B4.4.3 . . Metadata harvesting, metadata sharing
B4.4.4 . . Metadata ingestion
B4.4.5 . . Metadata update and maintenance
B4.4 . metadata attribution

B5 **Document processing**
All kinds of documents, text, speech, sound, still and moving images (see D). These methods by the system in building a collection and data store or on the fly by the user. NT B4.3.1
B5.1 . Document transformation
B5.2 . Automatic markup
B5.3 . Automated and computer-assisted knowledge extraction (fact extraction, relationship ex
B5.4 . Automatic translation
B5.5 . running linguistic and other analyses
B5.6 . extract citation links to make this one huge citation index, such as done in CiteSeer (http://citeseer.ist.psu.edu/)

B6 **Linking and content integration (especially large-scale)**

C **Facet C. Function: Digitization, preservation, access**

C1 **Collection development, selection, appraisal**

C1.1 . Rich content

C1.2 . Selection and appraisal

C2 **Digitization / OCR / digital formatting**

C2.1 . Digitization

C2.2 . OCR

C2.3 . Digital document structure XXX Architecture

C3 **Preservation and long-term access. Curation**

Provide systems and services that guarantee digital content will be accessible in the lon preservation of digital content requires urgently viable solutions.

Digital Preservation and curation -- the implications for the design of digital libraries th the need to ensure the long term access to the material they hold.

C3.1 . Preservation, persistence, digital longevity

C3.2 . Persistent identifiers

C3.3 . Long-term metadata viability

C3.4 . Selection and appraisal

C3.5 . Risk management approach to preservation

C3.6 . Assuring authenticity

C3.7 . Curation

C3.8 . preservation methods

C3.8.1 . . distribution of multiple copies

C3.8.2 . . Salvage and Rescue

C3.8.3 . . Preserve obsolete software needed to process old documents

C4 **Access and use**

C4.1 . **Content delivery, physical access,** obtaining a copy or otherwise being able to read the
C4.1.1 . . Integration of information (e.g., GIS combined with cultural heritage data, tourism)

C4.2 . **Intellectual access**. Search/Queries/Navigation
C4.2.1 . . Search engines
C4.2.2 . . Information retrieval. Enhanced retrieval functions. Search mechanisms
C4.2.3 . . Advanced query models
C4.2.4 . . Compound-object matching
C4.2.5 . . Navigational interface (beyond search)
C4.2.6 . . Result presentation
C4.2.7 . . Relevance feedback

C4.3 . **Working with material, individually and collaboratively**.
Application support/integration
C4.3.1 . . Annotation
C4.3.2 . . Document authoring. Users as contributors
C4.3.3 . . Provide the document processing functions from B5 to the user for ad-hoc applicat
C4.3.4 . . Collaboration among users

C4.4 . **Personalization / contextualization. Location-awareness / context-awareness in retr processing. Pervasive information**
Contextual information includes both explicit and implicit knowledge about end users, their environment. Such factors constrain the search without forcing the user to re-exp information need explicitly and frequently
C4.4.1 . . Personalization - adapting system behavior to persistent user characteristics
C4.4.2 . . Contextualization, context-awareness - adapting system behavior to the user's pres
Some context information collected dynamically by general environmental sensor
C4.4.2.1 . . . Location awareness - adapting system behavior to the user's present location)
C4.4.3 . . Pervasive information – context-aware, specifically location-aware, provision of in combining the system's general information with individual information about th collected dynamically by user-specific sensors.

C4.5 . **Mobile access**

C4.6 . **Digital library services**
C4.6.1 . . Digital library operational services / the pragmatic
C4.6.2 . . Advanced & specialized services
C4.6.2.1 . . . Knowledge-based services

C4.7 . **Reuse**

C4.8 . **User interfaces. Usability**
C4.8.1 . . User Interfaces
C4.8.2 . . Usability. Usability studies
C4.8.3 . . Accessibility for people with special needs

D **Facet D. Type of material**

D1 **Types of collections / systems by number of languages**

D1.1 . Monolingual collections / systems

D1.2 . Multilingual collections / systems
enable the user to access, interpret and utilize information of interest regardless of language boundaries
Note that diversity of language usually also means diversity of culture. This is important for retrieval and interpretation

D1.2.1 . . Multilingual systems, emphasis on language, natural language processing problems

D1.2.2 . . Multilingual systems, emphasis on cultural diversity, knowledge organization problems

D2 **Types of collections / systems by degree of control**

D2.1 . Uncontrolled collections (for example, the Web)

D2.2 . Controlled collections

D3 **Material by medium**

D3.1 . Plain text

D3.2 . Speech

D3.4 . Music

D3.5 . Multimedia - materials in all media, audiovisual objects

D3.6 . Solid (3-D) objects of all kinds (artifacts, museum objects, biological specimens)

D4 **Material by complexity**

D4.1 . Simple documents

D4.2 . Complex documents, compound documents, structured documents

D5 **Material by interactivity**

D5.1 . Non-interactive material

D5.2 . Interactive material

D6 **Material by static vs dynamic**

D6.1 . Static document

D6.2 . Dynamic material (changing over time)

D **Material by publication status**

D7.1 . Published materials (books, journals, distributed films, etc.) (many copies)

D7.2 . Semi-published materials (“grey literature”) (many copies, less widely available)

D7.3 . rare books, manuscripts (few copies)

D7.4 . unpublished materials, archival materials (often unique copies)

D8 **Material by copyright status**

D8.1 . Material out of copyright

D8.2 . Material in copyright

D9 **material by subject area**

D9.1 . Geographic information, geo-referenced content

D9.2 . Educational content

D9.3 . Cultural and scientific/scholarly content

Appendix 3. Inventory of additional ideas

This appendix is a list of ideas from the 2005 DELOS brainstorming meeting organized using the faceted classification given in Appendix 2. It could be the basis for an edited complete report from the meeting but is useful in itself.

This appendix can be found at:
http://www.delos.info/eventlist/BSM_Dec05/DELOSBrainstormingResultsOrganizedShortC.pdf

The following invited presentations were made at the brainstorming meeting:

The EU Vision - Horst Forster (EC)
http://www.delos.info/eventlist/BSM_Dec05/Horst_Forster_summary.pdf

The TEL Vision - Jill Cousins (Head of Office-The European Library)
http://www.delos.info/eventlist/BSM_Dec05/general_views_Jill_TheEuropeanLibraryVision.pdf

The DELOS Vision - Yannis Ioannidis (U. Athens, Greece)
http://www.delos.info/eventlist/BSM_Dec05/general_views_Yannis_niceVision051205.pdf

Overview of DL Activities in the US - Steve Griffin (NSF)
http://www.delos.info/eventlist/BSM_Dec05/Steve_Delos Dec05.pdf

Additional information about the meeting can be found here:
http://www.delos.info/eventlist/Brainst_dec05.html

Appendix 4.

Selected list of publications from the DELOS community

The following references provide a good view of the main research activities of the DELOS community. References 1 and 2 are the reports of two brainstorming meetings organized by DELOS to identify the research topics most relevant to Digital Libraries in the medium- and long-term. References 3 to 11 are the proceedings of a series of thematic workshops organized by DELOS to present and discuss the state of the art in a number of fields essential to the advancement of digital library technologies. References 12 to 20 are all contained in a special issue of the International Journal of Digital Libraries (August 2005) and represent the final output of a number of working groups jointly supported by DELOS and NSF.

A more exhaustive list of publications authored by the DELOS community at large can be found here:
http://www.delos.info/publications/DELOS_Publications.pdf

1. Digital Libraries: Future Directions for a European Research Programme, DELOS Brainstorming Report, San Cassiano, Italy, June 2001 http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf
2. Digital Libraries: Future Directions for a European Research Programme DELOS Brainstorming Report, Corvara, Italy, July 2004. http://www.delos.info/D8.1.2%20-%20Future%20Research%20Directions.pdf
3. Proceedings of the 1st DELOS Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, 11-12 December 2000. http://www.ercim.org/publication/ws-proceedings/DelNoe01/
4. Proceedings of the 2nd DELOS Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin City University, Ireland, 18-20 June 2001. http://www.ercim.org/publication/ws-proceedings/DelNoe02/index.html
5. Proceedings of the 3rd DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries, Darmstadt, Germany, 8-9 September 2001. http://www.ercim.org/publication/ws-proceedings/DelNoe03/index.html
6. Proceedings of the 4th DELOS Workshop on Evaluation of Digital Libraries: Testbeds, Measurements and Metrics, Budapest, Hungary. 6-7 June 2002. http://www.ercim.org/publication/ws-proceedings/DelNoe04.pdf
7. Proceedings of the 5th DELOS Workshop on Multimedia Contents in Digital Libraries, Chania, Crete, Greece, 2-3 June, 2003. http://www.music.tuc.gr/MCDL/
8. Proceedings of the 6th DELOS Workshop on Digital Library Architectures (Grid, P2P, and Service-Orientation), S. Margherita di Pula (Cagliari), Italy, 24-25 June, 2004.
9. Proceedings of the 7th DELOS Workshop on Audio-Visual Content and Information Visualization in Digital Libraries, Cortona, Italy,4-6 May, 2005.
10. Proceedings of the 8th DELOS Workshop on Future Digital Management Systems (System Architecture and Information Access), Schloss Dagstuhl, Germany, 29 March – 1

April, 2005.
http://www.delos.info/eventlist/delos-dagstuhl-handout-all.pdf

11. Proceedings of the 9th DELOS Workshop on Digital Repositories: Interoperability and Common Service, Heraklion, Crete, Greece, 11-13 May, 2005.
12. S. Griffin, C. Peters, C. Thanos, Towards the new-generation digital libraries: recommendations of the NSF/EU-DELOS working groups: Guest editor introduction, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 253-254. http://www.springerlink.com/(f1z34gni50bwxl55wxocih45)/app/home/issue.asp?referrer=parent&backto=journal,2,20;subject,73,150;
13. Y. Ioannidis, Digital Libraries at a Crossroads, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 255-265.
14. Y. Ioannidis et al., Digital Libraries Information-Technology Infrastructures, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 266-274.
15. C.-C. Chen, et al., Digital Imagery for Significant Cultural and Historical Materials, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 275-286.
16. J. Goldman, et al., Accessing the Spoken-Word, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 287-298.
17. A. Smeaton and J. Callan, Personalization and Recommender Systems in Digital Libraries, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 299-308.
18. G. Crane, et al., Emerging Language Technologies and the Rediscovery of the Past, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 309-316.
19. S. Ross and M. Hedstrom, Preservation Research and Sustainable Digital Libraries, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 317-324.
20. J. Borbinha, et al., Reference Models for Digital Libraries: Actors and Roles, Journal of Digital Libraries, Vol. 5, No. 4, August 2005, pp. 325-330.