A Research Agenda for Digital Libraries

Summary Report of the Series of Joint NSF-EU Working Groups on Future Directions for Digital Libraries Research

October 12, 1998

Editors: Peter Schäuble and Alan F. Smeaton

Executive Summary

Digital libraries represent a new infrastructure and environment that has been created by the marriage of computing, communications, and content, on a global scale. This supports individuals or organisations in a broad range of distributed knowledge based activities from electronic commerce to scientific collaboration. It is essential that this new type of functionality is developed in order to allow us to solve complex global challenges in areas such as business, environment, health, cultural heritage, and other important issues.

A collaborative effort among leading researchers from the US and Europe has been exploring the possibilities of a joint international research agenda in this field. Five working groups have been formed where each has explored a key research topic. These key research areas are Intellectual Property and Economics, Interoperability, Global Resource Discovery, Metadata, and Multilingual Information Access.

The working groups were assigned two co-leaders, one from Europe and one from the US and the groups have each met twice to identify the most important research directions and funding priorities. Full and detailed reports as well as coherent summaries of these research directions and recommendations have been prepared. These summaries are included in, and represent the main part of, this report which will be presented to US and EU funding agencies on October 12, 1998 in Brussels.

The recommendations by the five working groups contain a number of common threads.

- The first collective theme to emerge is increasing substantially the **level of collaboration** both across disciplines as well as across geographical boundaries. This is a reflection of the intrinsic nature of digital libraries which are multi-disciplinary and geographically unbounded. The summaries of the working groups list a variety of arguments for this important recommendation.

- A second general direction is the development of **new models and theories** in order to understand the complex interactions between the various components in a globally distributed digital library. This includes developing an understanding of the overall work tasks users face that bring them to use digital libraries. It is essential to understand all these interactions in a wider future context where components will interact which have not been combined in any way in the past.

- A third recommendation is to explore the development of **new information objects** and information genres, in particular digital-only objects having a highly dynamic nature which cannot be handled by traditional approaches.

The EU-US Joint Working Groups: Digital Library Project has been funded in part by the National Science Foundation under the US-Europe cooperative Science Program, REF: INT-9605202, and by the EU ESPRIT LTR programme under the DELOS WG (contract No. 21057).

- The fourth recommendation is to explore new functions to optimise **the handling of both new and traditional information objects**, in particular functions supporting new types of knowledge based activities.

- The next recommendation is to foster the **engineering, deployment and use of** large-scale, distributed and operational **digital libraries** which provide real content to real users. This has been shown to be a very fruitful approach in digital library work to date and the engineering and deployment should be continued.

- Finally, the groups recommend the setting up of a publicly available and re-usable **evaluation infrastructure** since the development of new digital library technology should go hand in hand with its evaluation. Evaluation in the context of a digital library means defining new metrics for components and new combinations of components in order to measure all sorts of performance aspects related to interoperability, metadata and resource discovery, the performance of the multilingual aspects of a digital library and the impact of intellectual property and economic issues.

The five working groups agreed that we are entering a terrain well beyond our past experience. The development of digital libraries is a huge challenge as well as a huge opportunity. There is much research to be carried out to develop the techniques needed, including long-term and applied research, as well as the development of infrastructures, standards, etc., in order to realise these challenges and opportunities. The recommendations reveal that future digital libraries require the integration of all sorts of components and aspects (software, methods, evaluation, content, etc.) that come from different disciplines as well as different geographic origins. Hence, we re-emphasise the importance of international co-operation and collaboration across disciplines and across nations.

Table of Contents

1. Introduction and Background

Costantino Thanos, IEI-CNR, Italy

Wide access to large information collections is of great potential importance in many aspects - economic, environmental, health, cultural, social, etc. - of everyday life. However, limitations in information and communication technologies have, so far, prevented the average person from taking much advantage of existing resources. There is no doubt that the quality of life could be greatly improved if people had easy access to the huge collections of information and knowledge that regard many facets of their lives. Indeed, humanity, in its continuous evolution, has accumulated an enormous quantity of information, knowledge, experience, art treasures, etc. One only has to think of the art treasures contained in our libraries and museums, or of the huge and precious collections of observational data in the areas of space exploration, earth sciences, the environment, medicine, etc., accumulated during the last century. A huge amount of material has also been produced by the entertainment industry (TV, movies, music).

A large part of these collections is currently available only on paper or in analogue form. Even if we ignore the problems of preservation, this fact poses severe limits on their accessibility as well as on the cost effectiveness of their management. Fortunately, recent advances in digital storage technologies for multimedia are making the digital archiving of vast amounts of information feasible.

However, even in the digital world, there are factors which prevent wide accessibility to very large digital collections. The main impediments are posed by i) the multimedia nature of these collections, and ii) the fact that current technologies do not allow an effective and efficient accessibility if the size of a collection increases considerably. Many models, techniques and approaches which work reasonably well with small and medium size text collections become inadequate as the collection grows. Models and techniques from databases, information retrieval, information filtering, document categorization and knowledge representation employed in modeling, accessing and managing huge multimedia information digital collections have shown severe limits in terms of effectiveness when applied to such kinds of collections. A number of important issues, such as content scalability, scalable semantic retrieval, etc. still remain open research problems. There is pressure from various sides, for instance from the emerging electronic commerce market, to solve these problems in the not too distant future.

We are conscious that, if digital libraries are to achieve their full potential, not only technological but also organizational problems must be addressed and solved. However, we mainly consider technological aspects here. Although organizational factors are clearly also crucial for the efficient and effective development of digital library systems, they are not the primary focus of this investigation, even though very important questions regarding intellectual property rights and economic issues have been examined. Also omitted from the brief of this joint NSF-EU collaboration are issues related to content provision which are recognised as being of equal importance to the technology being used.

In the US, the importance of digital library technologies has been recognized with the emergence of an advanced information infrastructure and the development of applications which have increased the connectivity of computer systems. After a preparatory phase, in which a number of workshops were held (see Source Book on Digital Libraries), in 1994 a Digital Library Initiative (DLI) was launched jointly by the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA). In the context of this initiative six large projects were funded. At the end of DLI, phase two was launched with the deadline for the first call for proposals in July 1998. An International Digital Libraries Cooperative Research Program has also been launched by the NSF with the deadline for proposal submission in October 1998.

In Europe, the digital libraries domain has not yet received the attention it deserves. However, some remarkable national initiatives, as well as some projects under the European Commission ESPRIT and Telematics programmes which address specific digital library technologies, show that also in Europe the importance of digital libraries is continuously gaining ground.

The Digital Library initiative of the European Research Consortium for Informatics and Mathematics (ERCIM) has a twofold objective: on the one hand, the aim has been to stimulate research activities in areas which are relevant for the efficient and cost effective development of digital library systems, to encourage collaboration between research teams working in the field of digital libraries and to establish links with on-going projects and activities in the field of digital libraries in industry and other public and private institutions, and, on the other hand, to contribute to the creation of a European digital library research community. As a leverage to be used to reach these objectives, two mechanisms have been adopted: the DELOS Working Group and a series of EuroConferences on digital libraries. The first has been funded by the ESPRIT Long Term Research programme and the latter by the Accompanying Measures of the TMR programme.

In carrying out these activities, it has been natural for European researchers to meet with the US DLI research community and very fruitful relationships have been established between these two research communities. We gratefully recognize and appreciate the support given by the NSF and the European Commission towards a closer collaboration between the DELOS WG and the DLI community. This collaboration has taken the form of five joint working groups on the topics of interoperability, metadata, intellectual property rights and economic issues, resource indexing and discovery in a globally distributed digital library, and multilingual information retrieval. Each working group has been composed of ten researchers, five from Europe and five from the US. To make European participation within these groups more representative, researchers from organizations outside the ERCIM consortium have also been involved.

The objective of these working groups has been to identify open research issues in the field of digital libraries and to make recommendations for future research actions. Each working group has organized the work according to its own needs. All groups have now met twice, once in Europe and once in the US. A mid-term meeting of all

working group leaders to coordinate and harmonize their work took place in June 1998 in Pittsburgh, USA.

Although the reports of the working groups are not finalized at the time of writing, we felt that it is our duty to provide our funding agencies with useful input for the definition of future programmes in the digital library domain. Indeed, now is a very important moment from this point of view. The definition of the specific research programmes of the fifth framework is under way; the second phase of DLI will start very shortly; preparation for a research program to be funded under the recently signed Agreement between the European Commission and the US for cooperation on Science and Technology is, also, in course.

Our ambition with the present initiative is to contribute to the definition of the future research programmes and projects of our funding agencies by recommending a set of specific research and development activities which we feel will most propitiously lead to advancement in the DL field. It is also our intention to improve and strengthen the existing cooperation between European and US researchers, and we feel strongly that this cooperation should be extended to other countries, given that global distribution is one of the main features of digital libraries.

We feel that the time is mature for a large digital library initiative in Europe. All the conditions in terms of scientific, application, industrial, and international contexts which call for and guarantee a successful research initiative, are currently satisfied. A European research community in the field of digital libraries has grown in an impressive way in the last few years as evidenced by the success of the European Conference on Digital Libraries in terms of the high scientific level and of the attendance of more than 500 participants. Additionally, the publication of the International Journal on Digital Libraries as well as other scientific publications summarised in Appendix I, testifies to the existence of a continuously growing European DL research community.

In many application/industrial areas (libraries, cultural heritage, health care, entertainment, protection of the environment, etc.) there is an increasing awareness that the building of very large heterogeneous digital information repositories, interconnected and accessible through global information infrastructures, requires further research in the digital library area. We have, thus, an application and industrial context that favors scientific and technological developments in the DL field. Finally, the European DL research community has now begun to collaborate with the US DL research community and, to some degree, also with the Japanese community. This workshop is a good demonstration of this fact. This means that a future European DL initiative will not be isolated, but will be well connected and integrated with similar US and Japanese initiatives. An international joint effort could promote relevant research actions along the lines that will be illustrated by the WG leaders in their presentations.

Such initiative should have two objectives:

- to foster and support long term research that addresses the many research issues of the field which still remain open;

- to fund a number of pilot projects aimed at developing large digital collections that can be used as testbeds for the validation of the research results. Indeed, an understanding of digital library issues requires operational experience that can only be gained by large scale deployment of digital library systems.

We are also convinced that DELOS has an important role to play in the future. First, it should continue in its promotion of activities aimed at creating a forum where a number of communities, i.e., library, cultural heritage, electronic publishing, electronic commerce, information infrastructure, software industry, etc., have the possibility of collaborating intensively with the DL research world, exchanging new ideas, experiences, and research results. Secondly, it should work towards the definition of a reference architecture for digital library services and encourage a number of pilot implementations of such an architecture. This activity should be conducted within an international cooperation and could lead to the definition of a standard DL architecture. Another important initiative that could be undertaken would be the development of large-scale, distributed testbeds, as experience has shown that progress on new digital library technologies depends on the availability of such testbeds. Finally, DELOS should promote the creation of working groups on particular topics when such topics are considered to be mature for a specific action. This essentially means following to some extent the operational model adopted by W3C. In order to be able to play such an ambitious role, DELOS must be transformed from a WG into a Network of Excellence. We feel that we are ready to undertake this effort.

This workshop is a first attempt to bring together researchers from the international DL community, representatives from US and European funding agencies, and European Union and national policy makers. The objective is to discuss concrete future research actions. We hope that our contribution will provide useful input and that a digital library initiative can be funded in the context of the 5th framework programme, as well as in the recently signed EU-US cooperation agreement.

2. Vision for Digital Libraries

Dan Atkins, University of Michigan, USA

The marriage of computing and communication on a global scale, combined with the increasing possibility of cost effective digitization and convergence of formerly separate media types has created the conditions for new infrastructure/environments to support humans (as individuals and organizations) in distributed knowledge-based activities. We are in fact in the early stages of understanding the implications of these technologies and even what to call the environments they will enable us to create. (We are at the "horse-less carriage" stage of understanding of form and function.) Terms such as "knowledge networks," "collaboratories," and digital libraries are being used with overlapping meaning. There is also a strong overlap between R&D on architecture for federating a distributed, autonomous set of digital collections and services, and the architectural requirements for large-scale electronic commerce. Some argue that digital libraries are in fact a specific case of an information economy (brokering environment.)

The concept of a digital library arose from the analogy with a place-based repository library containing an organized collection of print-on-paper and other physical artifacts combined with systems and services to facilitate physical, intellectual, and long-term access. The initial emphasis was on the retrospective conversion of print-on-paper objects into digital objects, usually page images, flat-text, compound documents of text and image, and/or structured documents. These digital versions of traditional library holdings offer distance-independent access, full-text searching, and potentially more powerful ways of finding timely and relevant information. Investment in retrospective conversion of physical objects to digital form, albeit not well coordinated, is continuing to increase.

Beyond retrospective conversion, the emergence of the world wide web has accelerated broad appreciation that we are now inventing new genres - new broadly understood document types - that have no print-on-paper equivalent but exist only in a digital computer/communication world. Furthermore wider audiences are now understanding that the underlying technology of digital libraries is so vastly different from the underlying technology of paper and ink. Digital information can be moved near the speed of light, stored at atomic scales of density, and converged into new document types combining, text, image, graphics, video, audio, hyperlinks, computational applets, and more. Digital libraries include the capabilities of physical libraries but potentially go well beyond them in scope and meaning. (What we are definitely not addressing well enough are questions of how we will preserve very long-term access to digital collections and continue to preserve our heritage with digital-only objects.)

One of the most obvious differences is the potential for a world wide digital library - not a centralized repository - but rather an enormous distributed but interoperable system or market of organizationally autonomous collections and services. Furthermore the line between the digital library as a system for access of information and the digital library as the environment for creating and sharing information,

particularly as part of a group activity will become increasingly blurred. The WWW is a crude, early form.

Most traditional libraries have already moved into the hybrid age of print-on-paper and digital resources but much more transformation is in the offing. One prominent researcher has proposed the following dimensions for exploring the potential differences between traditional and digital libraries:

- Traditional libraries are stable and slowly evolving; digital libraries are highly dynamic, ephemeral and versioned.

- Traditional libraries hold atomic objects of mostly print in big crisp chunks; digital libraries hold inter-linked, multi-media objects which are multi-size, fractal, and ill-defined.

- Traditional libraries hold objects with largely flat structure and minimal context and meta information; digital libraries support documents with significant internal scaffold structure and significant context/meta information which might be automatically extracted.

- Traditional (academic) libraries hold objects which are scholar-authored and pre-credentialled through a ponderous publishing stream; digital libraries allow anyone to publish in a lightweight way, and can support pre-credentialling or credentially through use.

- Traditional libraries are based upon centralized control and relatively few access locations; digital libraries can be distributed and ubiquitous.

- In traditional libraries the objects are physically and logically co-controlled; in digital libraries the physical and logical organizations can be separated (allows virtual collections).

- The tradition of public libraries is universal access and free; digital libraries could be similar in this regard, or digital libraries could support rich layers of access control and management of terms and conditions.

- Traditional libraries support one-way, loosely coupled (slow) interaction; digital libraries support two-way communication with tight, fast interaction.

- Traditional libraries are based upon a model of one-way search (a consumer looking for an object); digital libraries support symmetric search (consumer looking for an object an producer of the object looking for a consumer).

- In traditional libraries structured text queries (and some browsing) are used to aid intellectual access; in digital libraries complex interactions of query, navigation/browsing, and social filtering can be used.

This list is not necessarily good versus bad. It is rather an attempt to stretch the vision of differences between the digital library as a simulation of a traditional library and

entirely new modes of support the life cycle of information creation, distribution, use, and preservation.

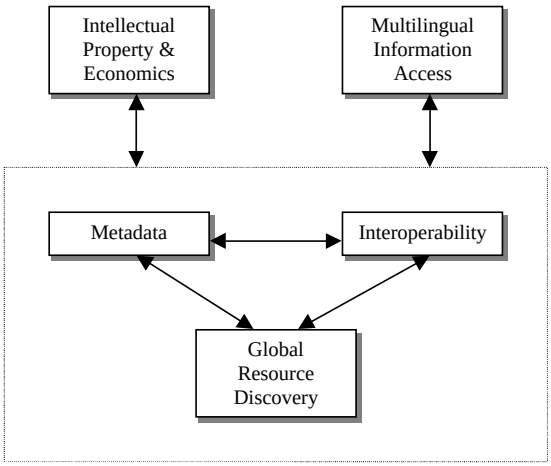
The realization of these potentials, particular in ways that are human-centered and serve all of society, is a major challenge requiring a complex interaction between social and technical disciplines, informed by theory, but grounded and informed by carefully selected pilot (testbed) projects. The NSF DLI 1 projects have stressed not only the need for relevant basic research in supporting technologies, but also for validating this research in the construction and field testing of real digital libraries with useful collections.

Our taxonomy for exploring a joint EU-US research agenda has been based upon five topical groups: interoperability, metadata, intellectual property rights and economic issues, resource indexing and discovery in a globally distributed digital library, and multilingual information retrieval. These 5 areas are an attempt to cover the key research areas relevant to the realization of a global digital library supporting access to millions of geographically and organizationally distributed collections and services. There will not be a central management authority, but rather an architecture which enables and provides incentive for producers and consumers of collections and services to find and transact with each other in mutually agreeable ways.

Interoperability, particularly in evolutionary ways, at many levels of abstraction, is a necessity and still poorly understood in a general way. Because of the scale of such an endeavor, metadata as surrogates and finding aids to target objects will continue to be of primary importance even as more structured, self-describing documents come into use. The issues of metadata standards, accommodation of heterogeneity in metadata, and automatic generation of metadata require additional work.

The greatest barrier to broader deployment of digital libraries with high quality, highly credentialled context are economic and intellectual property issues, appropriated re-interpreted and implemented in the digital world. Furthermore, economic models provide the basis for scarce resource allocation and creation of incentives for participation in the type of distributed, autonomous digital library mentioned earlier. The same technology that offers the potential to help people find information they need and want, is also, at least know, fueling information overload. This together with the dynamic, ephemeral nature of WWW-enabled digital libraries, places great demand on middleware to support intelligent resource discovered and characterization (e.g. indexing). And finally , multilingual issues, mostly ignored in the NSF funded DLI, is of critical importance in the EU and becoming more so in the USA. It is clearly critical to interoperability and the semantic level.

The way in which the 5 working groups interact can be summarised in the diagram below with the 3 areas of metadata, interoperability and resource discovery combining into a tight layer of enabling technology while issues of multilinguality and intellectual property and economics impact this group as a whole.



Different groups took differing methodologies in developing their contributions. For example, the resource discovery working group developed a vision and identified trends and used that as a reference for how they believe resource discovery should develop, whereas the working group on interoperability developed a reference architecture to be used as a framework for discussing research issues. Each is legitimate in its own context and was used to suit the needs and dynamics of the different groups.

3. Summary Reviews of the Working Groups

In the 5 sections which follow, for each of the groups a summary of the definitions of scope, the major research questions identified, the relationships to other working groups and the overall recommendations are presented.

A. Summary Review of the Working Group on Intellectual Property and Economics

Christos Nikolaou, University of Crete and ICS-FORTH, Greece.
Michael Wellman, University of Michigan, USA.

A.1 Scope of Inquiry

The future of digital libraries is marked by considerable uncertainty, much of which can be categorized as issues of intellectual property and economics. Many open questions exist, both about what is technologically possible, and what will actually happen. Various types of organizations–including libraries, schools and universities, publishers, learned societies, technology providers, as well as individuals (e.g., researchers, authors)–have a stake in the outcomes, and would benefit generally from a reduction in uncertainty. Research may shed light on these questions, by exploring possibilities, and by collecting and disseminating data about the current situation to inform choices made by the various entities that will shape this future.

A.2 Identification of Major Research Questions

Attempting to understand–much more so to design or influence–such a complex and dynamic system requires a broad array of tools from a variety of academic disciplines. The most directly relevant disciplines include social sciences (economics, psychology, sociology), information sciences (computer science, library science), and law. Although it is not possible to partition the research question by discipline–because many questions require multiple perspectives–our framework layers the concerns into realms loosely corresponding to these areas of study.

1. Institutional and social policy context. This includes specific legal questions such as interpretation of copyright in particular digital contexts, as well as broad questions about privacy policies.

2. Architecture and mechanism. This includes technological questions about how to construct computational infrastructure to support the information economy, and questions about the implications of deploying alternate mechanisms and policies.

3. Content and services. What information services are valued, how will they be used, what will they cost, and other types of questions about the content and offerings of digital libraries.

Any research agenda for DL.IPE must take into account the diversity of types of questions and stakeholders, and employ a corresponding variety of research modes and research disciplines.

Finally, we note that this research takes place within an extraordinarily dynamic environment, where new developments–technological, commercial, and institutional–regularly introduce dramatic structural changes to the operating environment. For this reason, it is doubtful that an agenda comprising specific questions to be addressed will be very stable. Nevertheless, examples of current questions are perhaps suggestive of possible near-term research achievements, and may provide a useful guide to those setting public research priorities.

A.2.1 Institutional and Social Policy Context

Any set of mechanisms that mediates exchanges among individuals and organizations is subject to some underlying institutional framework, including legal principles and statutes, mediating institutions, and conventions. For systems that deal with information goods and services, the most obviously relevant element of this underlying framework is intellectual property (IP) law. Applicable laws, and prevailing interpretations of those laws, define the status of existing intellectual property, and entitlements (e.g., fair use) of those "possessing" information. Moreover, IP law–in conjunction with broader commercial law–governs the forms of exchange of information services that are possible and enforceable within various legal jurisdictions.

It is commonplace to observe the incongruity of legal regimes crafted for old print media applied to new electronic realms. But rapidly changing technology leaves no alternatives; continual evolution of law and persistent uncertainty is to be expected as long as the target keeps moving. Interest groups exert political forces based on their best predictions of the effects of alternate regimes. One role of IP research is to inform the ongoing debates, so that all interested parties (including those who may not be actively participating in the process) may understand better the implications of current and proposed laws.

The policy context goes beyond IP law as well. Privacy policies dictate conventions or rules regarding dissemination of information about individuals. Certification authorities provide means to authenticate identity or other facts–for example providing pedigrees for digital documents. The broader legal regime and enforcement authority determines what rules must be respected, and what remedies are available when they are not.

A.2.2 Architecture and Mechanism

Given an underlying policy context, there will be a large space of possible system architectures, and component mechanisms, for constructing digital libraries. As noted above, the dynamic open nature of the library environment argues against a rigid architectural design. That is, any design must allow for new entrants, and accommodate new types of information goods and services.

Perhaps the most basic need is for information infrastructure specifically to support institutions mediating arrangements among entities participating in digital libraries. Taking the predominant type of arrangement to be economic, that is, based on exchange of goods and services, the major type of infrastructure required is that supporting commerce. This includes much more than payment mechanisms, indeed, payment is only a small part of the final stage in commerce: executing the exchange. A comprehensive commerce infrastructure would support the entire commerce cycle, including finding goods and services (and associated agents) of interest, negotiating terms, as well as the actual exchange. Components of such an infrastructure (i.e., the "middleware") might include:

1. Generic languages for describing information goods and services.
2. Advertisement and search facilities to match seekers and providers of services.
3. Endorsement services that provide or disseminate evaluations of other services.
4. Negotiation mechanisms facilitating agreement on terms of exchange.
5. Generic languages for contracts, including in particular intellectual property usage licenses.
6. Facilities for reasoning about licenses and other components of contracts.
7. Authentication mechanisms for individual identification, group membership, and license tokens.
8. Timestamping and "watermarking" mechanisms.
9. Other cryptographic services.
10. Exchange protocols, including digital payment.

This list is a rough cut, and there is no real need at this point to decide exactly what should or should not be included. Much of it is necessarily demand-driven, based on what research and experience at the service level suggests. That is, from the IPE perspective, the aim is to design architectures and mechanisms that support effective bottom-up organization of the overall digital library.

A.2.3 Content and Services

Computational infrastructure is but an empty shell for creation of digital libraries. The substance of what the library actually provides is in the available content and services. Indeed the content and service providers themselves, driven by demands of information consumers, represent perhaps the major source of innovation for digital libraries, present and future.

Those providing library services face difficult decisions regarding what services they should develop, how they should be delivered, and at what terms they should be offered. Those requiring services (perhaps including some of the same entities providing other services) face equally hard choices about which services to use, and what terms to offer. These are not necessarily different fundamentally from analogous choices in the non-digital or semi-digital realm. However, as the digital environment is newer and thus less familiar–and the multiplicity of choices perhaps greater–there exists a great need for guidance in how to approach these decisions.

For instance, many academic electronic journals are relatively new enterprises, run by research communities without special expertise in publishing. As far as we know, there exists no systematic compilation of guidance for those setting up new journals, advising them on the many critical strategic decisions they face. This is probably due to the fact that there is no accepted wisdom on such matters. Similarly, society publishers taking their collections online must blaze their own trails, as little collected experience yet exists on the successes or failures of such efforts.

Of course, there exists much work in the social sciences bearing on the behavior of researchers, learners, and other users of information. Similarly, economic models of consumers and producers may be applicable to these environments (though information goods have several special properties that can increase the complexity of analysis). However, these models require empirical calibration, often unsupportable by existing public data. For example, exactly what are the expected "first copy" costs of preparing content for various modes of distribution? The benefit side–that is, the value of information services–can be much harder to measure than the costs. Further research will be required to understand how to translate usage information in to more generally applicable measures of value.

A.3 Relationship to other Working Groups

Intellectual property and economic issues cut across the technological issues explored by the other working groups and need support from all aspects of the enabling technologies such as resource discovery, metadata and interoperability. Specifically, for resource discovery, issues of the usage of original data, copied data and cached data will be important as will indexes on data, metadata and registries. By its very nature, interoperability must support models which encompass intellectual property and economics issues.

A.4 Recommendations for Future Research

Although much useful research is being conducted and will continue by private entities, there exists a distinct public interest in DL.IPE research. Various ways that technology could develop will have different impacts on the respective stakeholders, and therefore there may be an important public sector role in facilitating those

developments deemed most beneficial. Understanding exactly what these effects are will itself be an important product of public-supported research.

Information about the economic environment (e.g., the demand for various forms of information goods and services) can be a significant strategic asset to participants in the information economy. For this reason, it is natural to expect that "market research" and other data collected by private entities may not be made available for public uses: decision making by policy makers and the broader class of participants. Therefore, it will be largely up to the public sector to provide such information, resulting in more principled policy decisions, and more effective deployment of resources by individuals, libraries, etc.

It is equally clear that the interest in DL.IPE research results spans national boundaries. Although many particular questions are specific to particular legal systems, existing academic institutions, or regional or cultural conventions, the information systems themselves will generally serve international constituencies. International coordination of research can serve both to make more effective use of research resources generally (organizing and disseminating results over a broader scope of participants), and to deal specifically with issues that arise at the interface of national systems (e.g., law, trade).

We recommend supporting a broad base of DL.IPE research, much of it multidisciplinary, focusing on issues with identifiable public interest and especially those addressing cross-national concerns. A successful body of effort must also include a combination of several research modes, including the following:

1. Fact finding. Several of the questions raised above hinge on relatively straightforward questions of fact. How are information systems used today, what do they cost, what does current law entail? We require further research on these and new questions, as well as means to facilitate collection and dissemination of these results.
2. Theoretical. The special properties of information goods call for new models and theories, and integration with existing models. Similarly, novel policy questions, introduction of large-scale computational infrastructure, and complex information use situations, for example, will all require extended theoretical developments to understand properly.
3. System building. Often the best way to understand a proposed mechanism, or to spur creation of new ones, is to build large-scale systems. To avoid premature standardization of whatever mechanisms are developed first, it can be advantageous to have prototypes of various kinds available for trials, and to spread the knowledge of what is possible.
4. Evaluation studies. An in-depth understanding of large-scale systems–either experimental or actually deployed–requires systematic purposeful evaluation. Novel theories similarly require empirical testing to judge their validity. For DL.IPE, a major question is what measures to use in evaluating systems and theories, and how to translate results into policy and design prescriptions.

B. Summary Review of the Working Group on Global Resource Discovery

Carl Lagoze, Cornell University, USA.
Norbert Fuhr, University of Dortmund, Germany.

B.1 Scope of Inquiry

Traditional information discovery and retrieval problems are compounded by distribution over the Internet. There is increasing volume, diversity, decentralization and autonomy in the development, meaning and types of information. The number of protocols for accessing this information increases and the reasons for making it available are more complex than simply sharing useful data. At the same time, there is a massive growth in the number and diversity of users' sophistication and background, and expectations. There is also an increasing criticality of the search problem to people's personal and professional lives.

Issues of quality and reliability are becoming more complicated. With increasing use of the Internet as a marketplace comes increasing incentives to abuse that marketplace with various forms of misinformation. The quality of the information will be questioned more and more as this trend continues. The increase in the diversity of sources of information is an additional complicating factor. Problems of context, provenance and timeliness become much more complex with the added dimension of distribution.

The solution to distributed information access will not be created by imposing a single monolithic solution on everybody. All solutions must be framed within organizational and economic contexts. The solutions must be targeted to support a world of different overlapping communities and permit layered solutions from no cooperation, to loose agreements, to tightly coupled organizations.

Group Methodology: The working group on Resource Discovery and Indexing used a number of vehicles to discovery important research areas critical to developing resource discovery tools for future digital libraries:

- *Vision* - We jointly developed a vision of how resource discovery in future digital libraries might occur. While the statement of the full vision is out of scope for this document, a brief summary is as follows. Interaction with existing global search services for the Web leads to the misconception that the future of distributed search simply involves improvements in quality of response to the list of keywords in a query. While improvements in precision and recall, for example, are important, efficient and effective distributed search will potentially enable the construction of entirely new classes of information-based applications. New kinds of information, new forms of user interaction, and new business models place entirely different demands on distributed search technology.

- *State of the art* - We compiled a short bibliography summarizing current work in this area. This bibliography will be published as part of the final report of the

group. To quickly summarize, current resource discovery research can be characterized as follows:

1. Attempts by information retrieval researchers (and web search engine vendors) to improve the effectiveness of centralized search engines.
2. Technologies to improve distributed indexing of multi-format documents: the Harvest project remains the most successful example of this.
3. Nascent research into protocols and techniques to effectively search distributed sources: this includes some of the early meta-searching efforts, content-based routing techniques such as GlOSS, protocols for meta-searching such as STARTS, and global digital library efforts such as Dienst.

- *Trends* - We enumerated a list of technical, social, and economic trends that we believe will provide a changed context for future resource discovery research. The final list of thirty items will be included in the final group report and includes such things as significant increase in connectivity and computing capacity, increase in the number of documents with access controls for security and pricing, and increase in the number of legal and regulatory restrictions to information access.

- *Problem Areas* - A categorization of the research topics related to resource discovery. These are described in the next section.

B.2 Identification of Major Research Questions

1. Systems Issues*:* Access to distributed information is hampered by the multiplicity of material available on-line from a network of public, private and commercial organizations, libraries, publishers, vendors and individuals. There is a great need for the development of a system infrastructure that facilitates navigation and retrieval, and that provides mediating support for the maze and variety of information available on-line. This system infrastructure should be capable of identifying, accessing, and retrieving the digital resources available. Furthermore, it needs to provide a coherent and consistent view of as many of the information repositories as possible. Three research threads are of central importance to the development of an architectural infrastructure that supports access to distributed information while ensuring acceptable behavior: query routing, database interaction, and consistency management.

2. Content Issues*:* A number of research questions related to content of resources were enumerated by the group:

- *Database Selection* - Metadata about the database content is required, as discussed in more detail in the working group on metadata, and this can be of different granularity, from frequency distributions of attribute values to high-level, condensed descriptions. Other factors affecting the selection are performance issues and pricing conditions. In order to perform an optimum selection, appropriate methods for deriving metadata of different granularity -- for all kinds

of media and representation languages -- and for estimating the relevant parameters have to be developed

- *Representations and query languages* - Future digital libraries will contain a variety of multimedia and hypermedia documents. For any type of document, there are three different views, namely the logical view (dealing with logical structure), the layout view (dealing with the presentation of the document) and the content view. Future query languages should include operators for specifying the logical structure, layout and content of the result. Also, order to support interoperability, standards for such query languages have to be devised.
- *Semantic heterogeneity* - A single database may contain a variety of document types, and different databases may be based on different schemas and use different query languages. From a user's point of view, many of these differences are not relevant for his information need. Thus, the system has to provide mechanisms for coping with semantic heterogeneity.
- *Ratings* - In order to guide users to the appropriate sources for his information need or for selecting the most suitable answers, ratings of databases or individual documents, which we refer to as metadata, will be essential for future digital libraries. Major criteria for ratings will be the quality of the material, the appropriateness for the current information need and filtering with respect to the user group (e.g. children). There is a need for implementing standards for the general structure and format of ratings such that a system can consider them during retrieval. For assigning the ratings, appropriate infrastructures have to be established.

3. Human-Computer Interface - Issues of HCI and distributed searching can be divided into four broad areas. The first two areas are the obvious ones of input and output. How do you get useful inputs from the users and how do you display the results? The other two broad areas are more cognitive and involve trying to understand what the user is trying to do and helping the user to understand what the system is actually doing and this should be applied to a complete spectrum of users.

- *Query Construction/Guidance* - The vast majority of searches consist of a single word. Such searches produce extremely large result sets. What can be done to help the user construct more focused searches? Some possible techniques include Query-By-Example, automatic query expansion and reference interviews.
- *Result Visualization/Presentation* - The majority of Internet searches result in ranked lists of documents. How can we improve on this? Better understanding of the content of the documents retrieved would allow more sophisticated presentation options. Exposing the indexes of the lower database layers could allow browsing of the indexes in addition to browsing documents.
- *Task Understanding* - Understanding of the task being performed by the user is the least mature of the areas of Human-Computer interaction being discussed here. Taxonomies of types of searches, types of results and types of searchers for a complete spectrum of user types, need to be developed. Understanding the user task can seriously alter how the search is performed.
- *Process Exposure/User Education* - Users are more productive when they understand how their tools work. Users have more confidence in the results of their searches when they understand how those results were produced. How can

we expose the distributed searching process to end users? Solutions in this problem area feed back directly into the issue of user provided guidance.

4. Organizational - In the field of distributed searching, every topic has an organizational aspect. Strategies for establishing collections and supporting technology always exist within an organizational context. The challenge is to organize large numbers of local networks, so that they can cooperate and take better advantage of their potential on a global scale, where the components are managed by many independent institutions with different goals and priorities and the systems range from state-of-the-art to dismal legacies.
Many aspects of distributed searching are technically difficult, even when the entire system follows the same technical standards and is managed to the same standards. Such issues include security, privacy, charging and quality of service. Because of these difficulties, it is tempting to seek for solutions that are effective in centrally controlled systems. Most such problems are challenging even when applied only in monolithic systems. The organizational problem is that almost all of these approaches lose their effectiveness without a substantial level of standardization and harmonization of management.

5. Research Facilities, Measurement, and Metrics - Like all areas of digital library research, the area of distributed resource discovery demands techniques to measure and evaluate the research. Three components need to be examined:

- *Testbed* - The evaluation of solutions for different distributed searching and indexing problems is quite chaotic at this point. Projects spend significant time building their own testbeds, using their own data, queries, and metrics. Furthermore, their results and conclusions are hard to compare with the results and conclusions of other related projects.
- *Taxonomies* - Researchers need to develop taxonomies describing different classes of users, user behaviors, queries, results, and roles. Part of the motivation for developing these taxonomies is that there is little known about the analysis of users' tasks and work practices requiring access to distributed information resources. Having these taxonomies will help define meaningful metrics and evaluate, for example, how systems deal with different query scenarios that might require different "kinds" of answers.
- *Simulation* - One extremely useful tool would be a set of simulations that build on the taxonomies and use the testbeds that have been developed.

B.3 Relationships to other Working Groups

Metadata - There is common focus on two areas. First, obviously, there is the very active area of metadata for resource discovery. The Dublin Core and RDF effort are examples of this. Second, there is the whole area of service metadata: for example, metadata published by a server that provides information on query routing.

Intellectual Property -There is overlap in the area both systems and organizational areas. In the systems area there is the question of access to metadata or surrogate data from content in order to allow indexing and resource discovery. In the organizational

area, there is the question of economic incentives and disincentives to encourage standards adherence.

Interoperability -There is overlap at the systems level, in the development of protocols to allow access to repositories and search sites.

Multilinguality - There is overlap in the user interface, systems, and content area. Obviously wherever metadata issues come up there is a question of multilinguality. In addition, presentation of results to users when the content is of mixed languages is an issue. In addition, there are overlaps in organizational levels, when standards must be developed at an international level.

B.4 Recommendations for Future Research

An active area within current digital libraries research is the standardization of the syntax and semantics of metadata as applicable to resource discovery and this will continue in the short term. A reasonable goal over the 2-3 year time frame is to formally characterize the distributed searching problem to include such issues as resource discovery and the characterization of network behavior that influences the performance of distributed search, and then develop formal metrics with which to measure success. A problem we struggled with throughout the working group meetings was developing a longer term vision of what distributed resource discovery really means - other than the "finding anything at any time" principle. A formal characterization of the problem including what users expect, how to help them with their tasks, etc. can do much to set priorities on the long-term research. Some work on this area has been done in the traditional library community vis-a-vis the "reference interview", but translation of this characterization to the digital library world has not yet been done.

As stated above, the remainder of our research suggestions are for the longer-term (5-10 year) time frame. Some of the most interesting and visionary work in this area will deal with resource discovery for multi-media documents and "dynamic" documents - those that change continually over time. We feel strongly that these types of document will play an increasing role in future digital libraries, and solutions for discovery of them are among the most challenging.

C. Summary Review of the Working Group on Interoperability

Hans-Jörg Schek, ETHZ, Switzerland.
Bill Birmingham, University of Michigan, USA.

C.1. Scope of Inquiry

The objective of the Interoperability Working group is to identify research issues related to the integration of different digital library (DL) sources and services. Whereas sources export documents or data of a certain format, services provide an interface to be used by other components or systems. Interoperability is important in the context of digital libraries: different DL sources and services have emerged and continue to emerge. These sources and services differ from each other in various dimensions. They may provide more or less different functionality, focusing on different aspects of digital libraries. It is a natural requirement that one wants to integrate some of these sources and services to a large system. There are many reasons, e.g., organizational issues or cost issues, why normally one cannot build a monolithic system with the same functionality from scratch. Consequently, research aiming at interoperability is fundamental in building large digital libraries.

The approach chosen by the working group was to define a reference or framework architecture that is depicted in Figure 1 which was developed purely in order to illustrate the major components and their interactions and to help identify research issues.

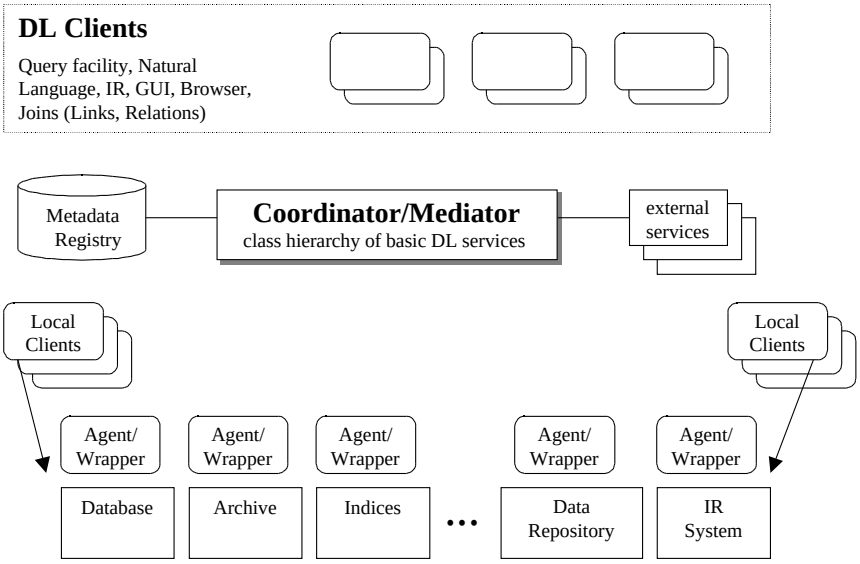


Figure 1: Reference architecture of an interoperable digital library system.

This architecture distinguishes between services and client applications. In between there is a layer responsible for coordination and control of combinations of services.

The following aspects regarding the reference architecture are important: as a rule, sources are autonomous, i.e., local clients may access the source and possibly alter its content without using the middleware architecture. As mentioned before, servers may export not only documents, but also services. Finally, we distinguish between data servers and external services. Essentially, whereas data servers provide storage services, external services offer other functionality. Based on this reference architecture, four main areas relevant for research have been identified.

C.2 Identification of Major Research Questions

The four main topics research regarding interoperability in digital libraries identified by the working group are the following ones:

- data/information model
- coordination and control
- query processing
- implementation mechanisms

The global interoperability layer for DL systems consists in fact of several interoperable DL services which abstract from a collection of heterogeneous and distributed data servers. These data sources are autonomous in the sense that local control over these sources is made. For example, the owners of the data sources are free to set conditions for their use (e.g., fees, usage rights), and are responsible for maintaining these sites. The data sources are integrated by agents that transform the specific access methods and local data models into the global information model. In addition to the basic DL services available in the mediator layer, external services might exist that have to be integrated. Relying on the mediator layer, global DL clients are able to query the interoperable digital library using sophisticated query languages like natural language, query-by-example or GUI browsers. The query facilities should include the capability to combine information from different data sources, like "joins" in database systems, either in *ad hoc* queries or materialized as links among different sources.

C.2.1 Data/Information model

The main goal of interoperability is to establish a global uniform view of the underlying data servers. Therefore an information model is needed which is able to express the structure and semantics of the integrated data as well as the available DL services. The description language based on this global information model should have an ontological basis for maximum flexibility and interoperability. Another area of research is the creation of document views on databases to facilitate information retrieval and vice versa.

Middleware Metadata

The middleware layer captures information about the underlying sources. This information is threefold: structure, semantics and services. Dynamicity is an important aspect of this layer, as the population of services and sites may be constantly changing.

A formal, ontological specification of the various information and service concepts and the relationships among them is needed. This is necessary to support automatic "finding" of services and information sources in a dynamic world. In such a world, the "names" of services and information sources may change or may be complex, requiring multi-faceted descriptions (e.g., an information provider may describe his service in numerous ways, such as by topics covered, recency of articles, and so forth). Furthermore, there may be various relationships among services and information providers that can be inferred through meta data (e.g., document X contains reprints of articles originally appearing in book Y, etc.). It is helpful, therefore, to represent metadata using multi-faceted descriptions and to provide automatic deduction (or induction) of useful relationships based on metadata. Formal ontological specifications of metadata naturally support both of these aspects.

Acquisition and exchange of metadata about the information sources' content and capabilities is important. It assists in the selection of sources relevant for a query, and the creation and optimization of queries against the source. Regardless of whether formal ontologies are used to represent metadata, it is important to have metadata of decent quality. Acquisition of high-quality metadata is a very difficult, yet critical task: without it, it is impossible to infer anything useful. Our point of view, in this regard, is to be fully compatible, and in fact, complementary, to existing metadata efforts, such as the Dublin Core.

Document Views

The classical understanding is that information-retrieval systems evaluate vague queries, whereas databases explicitly define the structure of the data. Database queries typically refer to an underlying schema. But one also wants to issue queries on database content that leave open certain characteristics of the underlying schema, or that are vague in some well-defined respect. A promising approach seems to be reducing the difference between databases and information-retrieval data and vice-versa: Creating database views on documents in information-retrieval systems. The problem to be addressed consists in a view definition mechanism to create meaningful term/weight information in an user defined context given by the objects and their relations. Such views must be indexed, the index structures must be kept consistent.

C.2.2 Coordination and control

The problem of providing coordination and control mechanisms for digital libraries has a variety of facets that future research should explore. As mentioned with regard

to information models, refinement of source-description languages is an important topic. Ideally, source description languages should be powerful enough so that a description of a source in this language allows for integration of new services into the infrastructure without implementation. Additionally, dynamicity aspects should be taken into account, i.e., how to cope with changes to the infrastructure or to the services, their interfaces and their functionality? Again, we would like to see solutions that do not require much implementation effort.

Furthermore, in a digital-library architecture, data administered by different servers may be modified. This may lead to consistency problems in case of replication between different servers. As opposed to self-organizing document collections, such as the WWW, digital libraries are subject to quality constraints and must incorporate solutions. The question that needs to be solved is how to provide transactional guarantees for different services if they are autonomous. For example, if a service is unavailable temporarily, this should not paralyze the whole library. There are solutions available for similar problems, e.g., spheres of atomicity or spheres of isolation, but it is still unclear how to adapt these concepts to the DL domain. A related problem is the efficiency of updates if update functionality of different sources is restricted. For example, a service may not allow deletion of index entries, instead the whole index has to be rebuilt. We would like to see solutions that compensate for this missing functionality in an efficient way.
The notion of quality, as mentioned above, is not only important with regard to data, but also with regard to query results. But, this notion has not yet been precisely defined. It would help not only to have an exact notion of result quality, but also have mechanisms for its visualization.

C.2.3 Query Processing

Query processing consists of several steps. Each of these steps imposes problems with a digital library assembled of heterogeneous components.

An important aspect of the problem of query processing in digital libraries is the formulation of queries for different sources. In principle, the "least common denominator" approach, i.e., provide only minimal functionality for declarative access, is not sufficient. Rather, the infrastructure for interoperability should compensate for limited declarative-access capabilities of individual sources. Already existing solutions to this problem cover only special cases, e.g., Boolean retrieval, but not more sophisticated ones, notably ranking of search results or search over structured documents (e.g., SGML-/XML-documents with different DTDs).

After query formulation, the query must be evaluated. In distributed environments, the query, or different sub-queries of the query, must be assigned to different sources, a task commonly referred to as query routing. Query routing is different in the homogeneous case, i.e., different sources are replicated throughout the system, and the heterogeneous case, e.g., WWW search engines return different results. Query routing depends on many factors, both logical ones, as well as physical ones, only some of which can be described using source-description languages. By and large,

current solutions only cover certain aspects of the query routing problem, but are not comprehensive.

Finally, relevance feedback has been shown to improve quality of query results, and respective mechanisms should be part of a digital library architecture. More generally, users can often assign value to a document, where relevance is one component of value. We need more general models of document value, such as those stressing pertinence. Such models can be very helpful for ranking retrieved documents from multiple sources, as well as improving the retrieval process.

The following research issues relate to interoperability in digital libraries and are relevant:

- How to design (and implement) mechanisms for non-topical relevance feedback?
- How to deal with feedback information even if different servers do not provide corresponding mechanisms?
- How to exploit feedback information with query routing? E.g., how to exploit feedback to decide which additional servers to address?
- General models of document (and service) value with subsequent investigation into the use of value for specific IR tasks, such as document ordering and retrieval.

C.2.4 Implementation mechanisms

At first sight, it seems worthwhile to implement architectures similar to the reference architecture using open standards. Given that there are various standards with high complexity, e.g., CORBA for distributed object technology, future research should not fall short of investigation of such standards. The development of standards for distributed systems in computing is a very volatile and fast-moving area, applicable not only to digital libraries but to other areas in distributed computing. It is important for digital library implementation mechanisms to track developments in this field and at the appropriate time to adapt best practice and the most fitting standards.

Interoperability implementation issues overlap with distributed computing issues such as network protocols and quality of service (QoS). Approaches to performance measurement will also be highly important.

C.3 Relationship to other Working Groups

We have already pointed out the different approaches taken by this working group and the working group on distributed search, and we have explained why both are legitimate in the particular context. Regarding the major research questions, there is a significant overlap with the ones from that working group, notably with regard to query processing/query routing.

Another close relation exists with the results of the working group on metadata. Interoperability in the DL context requires metadata. Consequently, the working group on metadata and the working group on interoperability have addressed similar

problems: source description languages, i.e., the aspects they address and their expressiveness, are a concern of both groups, as is the issue of translation between source descriptions.

C.4 Recommendations for Future Research

Interoperability is a central issue in the development and maintenance of digital libraries which by their very nature, will be distributed. Since digital libraries will be amalgamations of various services and information providers and since these amalgamations will be dynamic then effective, efficient and widespread interoperability is of paramount importance.

Further, as digital libraries are evolving with specific requirements different from self-organizing systems, such as the Web, it is important that research aimed specifically at digital-library interoperability be pursued. As interoperability mechanisms will form an infrastructure, rather than a variety of proprietary components, it is particularly appropriate that the research and development of this publicly usable infrastructure be undertaken with public rather than private resources. The 4 areas of models, coordination and control, query processing and implementation mechanisms provide an outline framework for this.

D. Summary Review of the Working Group on Metadata

Thomas Baker, GMD, Germany and Asian Institute of Technology, Thailand.
Clifford A. Lynch, CNI, USA.

D.1 Scope of Inquiry

"Metadata" is the Internet-age term for structured data about data. Typical examples are library catalog records, bibliographic headers in Web pages, and "terms of use" statements. Different user communities -- from librarians and computer scientists to government agencies, cultural heritage organizations, publishers, businesses, and the legal community -- scope and purpose metadata differently. As disciplinary communities such as biodiversity, earth and space sciences, and cultural heritage have begun to use networked information, they are developing specialized metadata systems. Networked information services will use metadata to interact with software agents and human users for a range of processes from resource discovery and document delivery to authentication, rights management and archiving. Metadata will also be an essential component in enabling electronic commerce and the shift to electronic publishing. The EU-NSF Working Group on Metadata sought to identify areas where further research is needed on architectures, tools, and models for managing metadata in the networked environment.

D.2 Identification of Major Research Questions

2.1. Models of metadata-resource association. Metadata can be embedded in resources, tightly coupled with resources using protocols or placed in separate databases. It can be created either by authors and content providers or by third parties; indeed, there is a large third-party metadata industry. The choice of model depends on technical, economic, and administrative considerations; we must understand how these choices affect performance and interoperability in large-scale distributed systems.

2.2. Metadata for service-mediated information. Proprietary digital content will increasingly be stored in systems that are closed off to uncontrolled indexing by Web harvesters. Similarly, many databases and legacy systems are present on the Web only through service interfaces and cannot be harvested. We need metadata that can formally specify services, policies, and transaction methods for such collections of information. Also, here, metadata must serve as a surrogate for the information itself, making it visible to searchers through indexing systems. Such mechanisms will be crucial for enabling large-scale distributed commercial publishing on the Web.

2.3. Metadata creation and management. Creators of metadata, who increasingly are non-specialists, need tools for generating, extracting, and managing metadata within efficient, automated workflows. In particular, metadata creation and management tools must become integrated with Web site management systems, database systems,

data warehouses, and other resources. Such production systems would improve the quality and cost-effectiveness of metadata in the networked environment.

2.4. Integration with information architecture standards. Metadata will promote interoperability to the extent that it can accurately be parsed and linked to reference schemas. The World Wide Web Consortium has proposed a Resource Description Framework (RDF) for expressing metadata in a way that is both usable by humans and processable by machines. The refinement of a generalized metadata architecture based on RDF will require several iterations as it is deployed for digital libraries and for commerce. ISO 11179, which addresses metadata registries and data elements, will need to undergo a similar evolution as it moves to practice.

2.5. Building registry systems. To make evolving standards effective, we will need to construct an infrastructure of registries that express schemas in both human- and machine-readable formats and offer authoritative guidelines for usage, local extensions, legal values, and mappings to other schemas. An ecology of registries may emerge that reflects a diversity of organizational motives, market forces, and user communities. There is a convergence of interest in such registries not just from the digital library community, for resource discovery, but from other service providers in government, business, and education and for a wide range of applications.

2.6. Core metadata sets. Current approaches to metadata, such as the Warwick Framework, think in terms of a series of relatively independent packages of metadata and how to exchange and use groups of these packages. The most mature metadata work is in description, with the Dublin Core package forming a central point of reference for descriptive practice. Similar core element sets are needed for other functions, such as structure and navigation, administration of digital objects, authentication, certification and provenance, terms and conditions, trust and quality, privacy, and longevity.

2.7. Interoperability and complexity. In all forms of human communication there is a tension between the need for simplicity and a desire for complexity. This has been apparent in the work on the Dublin Core descriptive metadata set. Such core schemas are extensible via additional elements or local refinements, but complexified adaptations can compromise interoperability. To control this, registries could use constructs such as interlinguas to link diverse ontologies of metadata among themselves. Work is needed on models and formalisms for describing such relationships and linkages among metadata schemes (mappings and crosswalks). Interoperability over time can be ensured only if such constructs are supported by social processes that allow user communities to negotiate global meanings while adapting them to local needs. As metadata sets are developed for functions other than description, we expect that these same issues will reemerge.

2.8. Metadata for complex digital resources. There are well-established conceptual models and practices for describing the contents of texts, their component parts, their formats, and their relation to other texts. However, we have limited understanding or consensus on the characteristics of the evolving digital genres -- resources that are more complex than normal text documents, such as collections of documents, dynamically-generated objects, and time-based media such as audio and video. As

these genres and our understanding of them advance, we will need appropriate metadata approaches.

2.9. Evaluation and metrics. As an RDF-based infrastructure for metadata is deployed, we have a unique opportunity to develop tracking systems to chart this deployment and to measure what types of metadata are being used within various communities. Research could investigate the costs and benefits of metadata, explore how discovery systems based on metadata compare with systems based on textual analysis and indexing, and determine which sorts of metadata deliver the greatest improvements in performance.

2.10. Policy issues. Metadata will be used across national and cultural borders; its interpretation and legal standing may depend on these. We cannot trust metadata to be accurate unless we know and trust its source; this is unacceptable barrier to commerce and knowledge sharing on the Internet. This is a problem on the Web today; for example, text-based Web indexers have had to work hard to compensate for "index spamming," whereby authors try to make their documents appear relevant when they are not. Unless research extends this to metadata, resource discovery systems will be limited to using metadata from a few big trusted sources.

2.11 Metadata diversity and resource discovery. While we have much experience with homogeneous descriptive metadata (i.e., the library catalog), very little is known about how to effectively re-purpose and integrate the much broader range of metadata now being associated with digital objects into the discovery process. The availability of various types of metadata will be highly variable in practice. We do not know how to design discovery and retrieval systems that make good use of all of those types, some of which will be only sparsely available. Enhances in discovery motivate the deployment of metadata, while investments in advanced discovery systems will be motivated by the availability of metadata to support them. Thus, research progress in resource discovery will also produce incentives to deploy rich metadata.

D.3. Relationship to other Working Groups

Of most relevance to the Working Groups on Interoperability and the Working Group on Resource Indexing and Discovery are models of metadata-resource association (2.1), metadata for service-mediated information (2.2), integration with information architecture standards (2.4), and building registry systems (2.5). Diverse metadata and resource discovery (2.11) is at the intersection of the work of our group and the group on Resource Indexing and Discovery.

Of most relevance to the Working Group on Intellectual Property and Economic Issues are metadata for service-mediated information (2.2), core metadata sets (2.6), metadata for complex digital resources (2.8), evaluation and metrics (2.9), and policy issues (2.10).

Of most relevance to the working group on Multilingual Information Access are building registry systems (2.5), core metadata sets (2.6), and interoperability and complexity (2.7).

D.4 Recommendations for Future Research

Dublin Core and RDF seem poised to provide a metadata system that is consistent across a wide range of applications and domains, usable by both experts and non-experts, interoperable with existing library catalogs and legacy databases, and coherent across many languages.

Near and Medium term activities

In general, the research topics outlined above focus on issues that will be central to fostering the growth of networked information resources, digital libraries, electronic commerce, and network-based publishing. A commitment to work on registry infrastructures could provide an important focal point and source of cohesion for research and development in many of the areas we have identified, as well as being an important research project in its own right.

Effective research progress on metadata needs to involve intense collaboration between metadata specialists and communities trying to solve functional problems, such as rights management, resource discovery, or archiving and preservation.

Research questions for the long term

The definition and management of metadata over time is a complex social process requiring negotiation, consensus-building, and iteration. Learning to manage these processes effectively and to coordinate the ever-growing activities of many disparate communities of interest is clearly a long-term research undertaking involving complex economic, technical, and social questions.
We have no experience with something of this complexity and scale and it is important to have some continued investment in theoretical and foundational work that will help us to deal with this evolving complexity.

E. Summary Review of the Working Group on Multilingual Information Access

Peter Schäuble, ETHZ, Switzerland.
Judith Klavans, Columbia University, USA.

E.1 Scope of Inquiry

The world is becoming more interconnected every day. The World Wide Web, electronic mail, distance collaboration, digital libraries, electronic commerce, and an increasing number of similar capabilities are making all kinds of information globally available, not just in English but in many other languages. Today's users of the international information networks come from industry, commerce, government, research, medicine, law, and indeed from all fields of life. Although these users may have widely varying degrees of language skills and many may have little or no expertise in more than one language, they must be able to access pertinent information in whatever language it appears. The growing user community is thus creating enormous pressure for access to information without language or cultural barriers.

The goal of the Working Group on Multilingual Information Access (MLIA) was to target short and long term research questions that are critical to the development of future multilingual capabilities for information systems. We have focused on major problems requiring solutions in the multilingual area since other information access issues are already addressed by related working groups on global resource discovery, metadata and interoperability.

Questions addressed by the Working Group on Multilingual Information Access can be grouped into two sets, involving:

1. Data Exchange - includes issues such as character encoding, font displays, browsing, etc. Such issues have implications for the international sharing of both original data and metadata as they must be interpreted, parsed, and displayed by Web-browsers, search engines, transliteration, transcription, and other systems.
2. Language Processing - covers natural language processing technologies, e.g. syntactic or semantic analysis, machine translation, information retrieval or information discovery in multiple languages, cross-language access, speech processing, and summarization. Multilingual language resources, such as dictionaries and thesauri, corpora and test collections, are also considered.

MLIA is a cross-community interdisciplinary enterprise and involves, at the very least, the following areas of research:

- Information Retrieval, which includes free text indexing, query processing, retrieval mechanisms, creation of metadata, etc.
- Machine Translation, which includes transfer and interlingual techniques, etc.

- Computational Linguistics, which includes morphological analysis, parsing, term recognition and expansion, semantic analysis, disambiguation, language generation, and so on.
- Document Processing, which includes clustering and classification, filtering, document segmentation, etc.
- Resource Development, which includes the construction and maintenance of dictionaries and thesauri, index terms, terminology, creation of derivative lexicons, domain-dependent lexicons, ontologies, building of collections for testing and evaluation, etc.
- Human Computer Interaction/Presentation, which includes visualization of document sets, single- and multi-document summarization, etc.
- Speech processing, which includes speech analysis, recognition, and generation.

In addition, results from the following fields are relevant to MLIA:

- Character-level data management, including encoding, display of multiple fonts and languages, transliteration, etc.
- Databases, including interaction with distributed heterogeneous databases, multilingual metadata such as that proposed by the Dublin Core, etc.

The relative importance of these areas to MLIA will be discussed in the full report of the working group and summary recommendations will now be made as to those areas in which future research should be concentrated.

E.2 Identification of Major Research Questions

The research problems can be grouped into three major areas: user needs, technology, and resources.

E.2.1 Definition of User Needs in Multlingual Information Access

A major challenge is to build the necessary infrastructure to study how users interact with multilingual information, and what their specific needs are for reaching across language barriers, both for access to information and for communication around access. This is essential in order to evaluate approaches to truly global, multilingual, and multicultural information access. Attention should be given to studying the following user requirements:

- the accommodation of variance in users within the multilingual situation: language comprehension and language manipulation.
- the capture of user feedback in MLIA to
    1. improve queries,
    2. improve result presentation,
    3. improve translation, and
    4. test new MLIA interfaces for evaluation.

A small set of initial basic evaluation tools are available primarily from United States Government Agencies including the National Institute of Standards (NIST) and the Defense Advanced Research Projects Association (DARPA) which funded the Text REtrieval Conference (TREC), with collaboration from European groups for cross-language system assessment. In some cases, these tools were created for narrowly construed tasks, and can be adapted to other evaluation tasks. However, what is lacking is a more general infrastructure to evaluate approaches to information access in a wider international context for a broader set of information seeking situations.

E.2.2 New Technology Research Areas

Problems unique to multilingual information access require solutions to a comprehensive set of issues including that of representing documents in different languages, correlating representations and terminology in different languages, accurate retrieval without necessitating exact query translation, and coherent presentation of linguistically diverse material. The technical challenges in MLIA go beyond a simple coupling of standard monolingual information retrieval (IR) techniques with still-developing machine translation (MT) systems. Indeed, one of the major misinterpretations of the technical challenges in MLIA concerns the relationship between Information Retrieval and Machine Translation engines since MLIA is not just a pipeline of IR + MT + Resources.

Specific areas where research can be usefully focused are:

- multilingual indexing tools - no fully multilingual indexing tools exist as yet and research is required on the most effective way to achieve this goal;
- user queries; the multilingual aspects of query interpretation and expansion should be studied;
- document clustering - for efficient access, multiple documents should be organized into classes that allow user to examine only most relevant documents; procedures for clustering over languages should be studied, as should cross-language document matching techniques;
- summarization - the implications of summarizing the results of a multilingual search (i.e. presenting all results in the query language) should be studied as should the possibility of merging information from documents, even when they are in several languages;
- visualization tools - so far few explicitly multilingual visualisation tools have been developed; more work is needed in this area;
- multilingual multimodal and multimedia access - the integration of standard text-based access across modalities (e.g. speech input and output) and across media types (e.g. video, sound) raises a set of specific technical challenges in multilingual information access.

E.2.3 Resources Required for Technology Development and Evaluation

A thorough set of well-developed resources is essential for achieving useful systems, both in terms of the development of new techniques and methods and for careful

evaluation. Much work has already been done with respect to construction of language-oriented resources in previously funded programs both in the European Union and in the United States, but what is freely available at this point is deficient in many respects. Research efforts should be concentrated on the following areas:

- the establishment of standard collections;
- the extensibility of resources, e.g. to include more text or additional capabilities; merging of existing resources. Attention should be paid to development of semi-automatic and automatic updating procedures since the goal is for new systems to have the capability to acquire new data automatically;
- the extension of resources to cover new languages. This includes the extension of existing monolingual resources to multiple languages, and monolingual processing to multilingual processing;
- the identification and development of language independent resources, such as ontologies and conceptual thesauri;
- the building of truly multilingual, rather than just bilingual, language resources;
- the creation of standards and evaluation procedures by which to measure new resources; this includes the building up of multilingual test collections;
- the understanding of the relationship between linguistic resources and MLIA effectiveness.

E.3 Relationship to Other Working Groups

Multilingual Information Access needs support from techniques evolving in three areas, each covered by another working group in this program. Most system, content, and Human Computer Interaction (HCI) issues discussed by the Global Resource Discovery working group are directly related to MLIA. The availability of Metadata will improve in general the accessibility of information, not only multilingual information. Indeed, some of the metadata values may well be language indices directly resulting from some of the language techniques outlined above, e.g. document profiles, multilingual terminology, etc. Similarly, the availability of structural information as discussed by the Interoperability working group will improve in general the accessibility of information. Finally, legal issues discussed by the Intellectual Property and Economics working group also affect the general accessibility of information.

E.4 Recommendations for Future Research

The recommendations of this working group are to study, design, and build the necessary infrastructures to analyse the user needs and to evaluate systems which provide multilingual, and multicultural information access. As such, this requires integration between groups from diverse language and cultural backgrounds.

We also recommend that the different research communities listed above be encouraged to combine their technologies and language processing skills to achieve a common interdisciplinary goal. Technologies must be developed capable of integrating and exploiting existing resources and procedures in order to achieve

efficient and effective fully operational MLIA systems. Standardization is very important for both static (language-oriented) and dynamic (procedural) resources. Much work has already been done in the area of lexicon and resource formatting; attention should now be focused on developing standards for the acquisition and use of dynamic resources. Focus should be on moving beyond existing static bilingual resources to fully multilingual dynamic resource acquisition and exploitation.

Internationally linked funding programs are highly recommended to bring different communities together along two dimensions: first, communities with different language and cultural backgrounds and second, communities covering different research areas that had little interaction in the past. The members of the MLIA working group agree that they have benefited from this multi-dimensional mixture when preparing this report.

4. Summary of the Recommendations

There are a huge set of challenges in mastering the size and the complexity of digital library systems which have a place in electronic business, in disseminating information on cultural heritage, science and technology, environment, health, and so on. The increase in size and complexity, in comparison to what has been achieved to date, will by itself be a major challenge of the future.

The brief of this joint NSF-EU initiative was to explore the **technical** issues only. A digital library deployed in an operational context will be global and will be distributed and will encompass a range of content and services, and a large spread of users. However with only a technological framework, a digital library is nothing. It is by combining the technology with real content that we get a true digital library. It should be understood that by concentrating on exploring only the technical issues as we have done in this summary report, we have omitted similar explorations into content provision. Whatever successes the limited number of digital library projects have had to date in the United States, Europe and Japan, they have all been based on operational and deployed digital library systems where access to real content which is meaningful to its users has been provided.

A digital library is the integration of multiple components which do not initially fit together in a seamless fashion for a number of reasons. Firstly, the necessary components come from a background of different communities and secondly, they should – in combination with other components – enable new functions which were not under consideration when the individual single components were designed and implemented. This means that the realisation of large-scale globally distributed digital libraries depends as much on collaborative effort as it does on the development of new technologies in order to develop systems which truly integrate their components. The level of collaboration required, across disciplines as well as across geographical boundaries, will be much higher than we have previously encountered.

In trying to bring together the respective summaries of the 5 working groups we have identified the following common threads:

1. We should avoid reinventing individual technologies or "wheels" by promoting cooperation among groups that have already developed a component technology or "wheel" which is ready to be used by another group, such as multilingual information retrieval and linguistic resources or metadata and resource discovery. Part of this task also involves discovering the component technologies that need to be or can be combined. The cooperation of groups from different communities requires the consolidation of their domain specific vocabulary which is a prerequisite for making progress.

2. The desire for new functions and the complexity of combining old and new functions require new theories and models so that we can understand the interactions between these complex components in a wider context. This, in turn, is a prerequisite for developing the new generation of digital library systems.

3. In order to explore the issues of size and complexity which appear throughout digital libraries, the most successful approach is one based on engineering large-scale, distributed and operational digital libraries which provide real content to real users. The emphasis here should be on practical applications and this means building, deploying and then using and evaluating aspects of such systems such as their efficiency, effectiveness and usability. The motivation behind this is to get people to use digital libraries and also to foster content providers to create and contribute to such systems.

4. Evaluation in the context of a digital library means defining new metrics for components and new combinations of components in order to measure all sorts of performance aspects related to interoperability, metadata and resource discovery, the performance of the multilingual aspects of a digital library and the impact of intellectual property and economic issues. The over-arching rationale for doing this is to build publicly available evaluation infrastructures which can be re-used.

5. We can observe that, in the early days, the web benefited from the flexibility with which almost anybody could create and publish content, but as it currently stands the web is now paying the price for this flexibility because access to relevant material is such a difficult process. Activities related to metadata aim at swinging the pendulum back towards improving the accessibility to information in a digital library by increasing the degree of standardization and conformance, while at the same time allowing the flexibility which has encouraged creation of content to continue.

6. One of our most significant conclusions from the work which has gone into preparing this report is to re-confirm the importance of international cooperation and collaboration. The need to bring together research communities which heretofore have had little overlap is the first justification. A second justification is based on the fact that digital libraries are inherently distributed, which is part of their appeal, and it is crucial to share aspects such as standards to ensure that the functions a digital library provides are useful in a global, multi-lingual and multi-cultural environment. A final justification is related to content in that the contents of a digital library should be material whose appeal is not limited to geographic or national boundaries.

The above trends have been identified from the contributions of the 5 individual working groups but further details can be found in the earlier sections summarising the groups' activities and in the final white paper on this NSF-EU initiative which will be completed shortly.

Appendix I: Sources of Further Information

The challenges of researching and realising Digital Libraries is now assuming its own place in the scientific community through specialist conferences, journals and publications. The following is an indicative list of the most important specialist publications in addition to the many other forums in areas related to digital libraries which have started to accommodate the presentation of such work.

Journal Special Issues on Digital Libraries:

- Communications of the ACM. Special Issue on Digital Libraries. April 1995, 38(4).
- ERCIM (European Research Council for Informatics and Mathematics) News. A Special Issue on Digital Libraries. October 1996, Vol. 27.
- IEEE Computer. Special Issue on the U.S. Digital Library Initiative. May 1996.
- IEEE Computer. Special Issue on Digital Libraries Challenges scheduled for 1999.
- SIGLINK Newsletter. Special Issue on Digital Libraries. September 1995, 4(2).

Digital Library Journals:

- D-Lib Magazine. "The Magazine of Digital Library Research." July 1995 - Present.
- International Journal on Digital Libraries. Springer-Verlag: Berlin, Germany. 1996-present.

Conferences/Proceedings:

- Organizing the Global Digital Library (OGDL): Theory and Practice of Digital Libraries.
- The International Symposium on Research, Development, & Practice in Digital Libraries (ISDL).
- ACM International Conference on Digital Libraries.
- International Conference on Conceptions of Library and Information Science (CoLIS).
- IEEE Advances in Digital Libraries (ADL) Conference.
- European Conference on Research and Advanced Technology for Digital Libraries (ECDL).
- DELOS Series of Workshop Proceedings.

Appendix II: Working Group Members

Co-organizors: Dan Atkins, University of Michigan, USA
Costantino Thanos, IEI-CNR, Italy

Intellectual Property and Economics:

Costis Dallas, Greek Ministry of Foreign Affairs
P. Bernt Hugenholtz, University of Amsterdam, The Netherlands
Jeffrey MacKie-Mason, University of Michigan, USA
Christos Nikolaou, University of Crete and FORTH, Greece
Ann Okerson, Yale University, USA
Bernard Rous, ACM, USA
Jakka Sairamesh, IBM T.J. Watson Research Center, USA
Pamela Samuelson, University of California, Berkeley, USA
Sebastien Steinmetz, Econometrics Lab of Ecole Polytechnique
Christine Vanoirbeek, ERCIM, Switzerland
Michael Wellman, University of Michigan, USA

Interoperability:

William Birmingham, University of Michigan, USA
Klemens Boehm, ETH, Switzerland
Sophie Cluet, INRIA, France
Panos Constantoupoulos, FORTH, Greece
Vassilis Christophides, FORTH, Greece
Barry Leiner, MCC, USA
Olle Olsson, SICS, Sweden
Andreas Paepcke, Stanford University, USA
Fausto Rabitti, CNR, Italy
Hans-Joerg Schek, ETH, Switzerland

Metadata:

Gene Alloway, University of Michigan, USA
Thomas Baker, GMD, Germany, and AIT, Thailand
Howard Besser, University of California, Berkeley, USA
Jose Borbinha, INESC, Portugal
Rachel Heery, UKOLN, UK
Ole Husby BIBSYS/SINTEF, Norway
Renato Iannella, Distributed Systems Technology Center, Australia
Carl Lagoze, Cornell University, USA
Clifford A. Lynch, CNI, USA
Shigeo Sugimoto, University of Library and Information Science, Japan
Anne-Marie Vercoustre, INRIA, France
Stuart Weibel, OCLC, USA

Multilingual Information Access:

Jaime Carbonell, Carnegie Mellon University, USA
Bruce Croft, University of Massechusetts, USA
Vasilieos Hatzivassiloglou, Columbia University, USA
Eduard Hovy, University of Southern California, USA
David Hull, Rank Xerox Research Centre, France
Judith Klavans, Columbia University, USA
Doug Oard, University of Maryland, USA
Carol Peters, IEI-CNR, Italy
Peter Schauble, ETH, Switzerland
Paraic Sheridan, ETH, Switzerland
Gary Strong, NSF, USA
Shigeo Sugimoto, University of Library and Information Science, Japan
Evelyne Tzoukermann, Lucent Technologies, USA

Resource Discovery in a Globally Distributed Digital Library:

Bill Arms, CNRI, USA
Mic Bowman, Trans-Arc, USA
Norbert Fuhr, University of Dortmund, Germany
Luis Gravano, Columbia University, USA
Sarantos Kapidakis, University of Crete, Greece
Laszlo Kovacs, MTA-SZTAKI, Hungary
Carl Lagoze, Cornell University, USA
Ralph LeVan, OCLC
Mike Papazoglou, Tilburg University, The Netherlands
Alan Smeaton, Dublin City University, Ireland