

**INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME**



Contract for:

Shared-cost RTD

Annex 1 - "Description of Work"

Project acronym: MIND

Project full title: Resource Selection and Data Fusion for Multimedia
International Digital Libraries

Proposal/Contract no.: IST-2000-26061, MIND

Related to other Contract no.):

Date of preparation of Annex 1: 03 June 2003

Operative commencement date of contract: 1 January 2001

1	PROJECT SUMMARY	3
2	PROJECT OBJECTIVES	3
3	PARTICIPANT LIST	4
4	CONTRIBUTIONS TO PROGRAMME/KEY ACTION OBJECTIVES	4
5	INNOVATION	5
6	COMMUNITY ADDED VALUE AND CONTRIBUTIONS TO EU POLICIES	6
7	CONTRIBUTIONS TO COMMUNITY SOCIAL OBJECTIVES	6
8	ECONOMIC DEVELOPMENT AND S&T PROSPECTS	6
9	WORKPLAN	8
9.1	GENERAL DESCRIPTION.....	8
9.2	WORKPACKAGE LIST	11
9.3	WORKPACKAGE DESCRIPTIONS.....	12
	WP1: ARCHITECTURE, CONTENT AND CONTEXT (UNIDO).....	12
	WP2: RESOURCE DESCRIPTIONS (DSI)	13
	WP3: RESOURCE SELECTION (UNIDO)	15
	WP4: DATA FUSION (USFD).....	17
	WP5: HETEROGENEITY (USFD).....	19
	WP6: EVALUATION (USG).....	20
	WP7: DISSEMINATION (USG)	22
	WP8: EXPLOITATION (USG)	23
	WP9: PROJECT MANAGEMENT (USG)	24
9.4	DELIVERABLES LIST	26
9.5	PROJECT PLANNING AND TIMETABLE.....	28
9.6	GRAPHICAL PRESENTATION OF PROJECT COMPONENTS.....	28
9.7	PROJECT MANAGEMENT	28
10	CLUSTERING	30
11	OTHER CONTRACTUAL CONDITIONS	30

1 Project summary

MIND addresses issues that arise when people have routine access to thousands of heterogeneous and distributed multimedia Digital Libraries. When so many Digital Libraries are available, the first information access task is *resource selection*. This is predominantly an ineffective manual task as users are unaware of the contents of each individual library in terms of quantity, quality, information type, provenance and likely relevance. People need accurate automatic resource selection tools to assist them. Once a set of libraries is selected and searched, a person must organise and interpret the (possibly multimedia) information supplied by different Digital Libraries. Typically this is performed through visual evaluation and ad hoc integration which forces users to restrict their attention to a small subset of the information retrieved. As the number of Digital Libraries increases the problem is exacerbated and the aspirations of information providers to increase access to scarce and/or unique resources is hindered. People need user interfaces that enable them to more fully exploit the multimedia information they find.

MIND will assist users to know where to search, how to query different media, and how to combine information from diverse sources. Therefore, the key objective of the MIND project is to address the problems faced by users in terms of their ability to access and exploit the increasing number of Digital Libraries available internationally through networks, like the Internet and the world wide web (WWW). More specifically the objective is to design models and to build sets of tools and associated test-beds to improve the effectiveness of resource selection, multimedia information access, retrieval and fusion of the retrieved data. The achievement of this objective will involve:

- ◆ the development of a variety of metadata generation methods for different media;
- ◆ the design of algorithms for improving the selection of the most relevant collections of information;
- ◆ the development of data fusion techniques for merging ranked lists of items retrieved from the selected collections;
- ◆ the evaluation of these methods on different sources of data and cohorts of users.

The achievement of this objective will be evaluated at different stages of the project using standard effectiveness measures (e.g., precision and recall) and user-studies. The quality of the research work carried out will be evaluated through presentations at refereed international conferences and publications of peer-reviewed journals.

2 Project objectives

Research groups worldwide are directing significant effort towards the creation of sophisticated Digital Libraries in a variety of disciplines. At the same time, a large number of companies, government agencies, and individuals are creating simpler Digital Libraries that are accessible from corporate networks or the Internet. Today, a person must know *where* to search, *how* to query different media, and how to *combine* information from diverse resources. As Digital Libraries continue to proliferate, in a variety of media, and from a variety of sources, these problems of *resource selection* and *data fusion* become major obstacles. The emergence of multimedia, including text, recorded speech and images, only exacerbates current problems and emphasises the need for new solutions. An answer to a query might be in a text document, an audio clip, or a newswire photograph. Effective, *reliable* information retrieval requires the ability to pose multimedia queries across many Digital Libraries.

The information access environment that we envision requires advances in several different areas. The proposed research is an *end-to-end* solution, covering how Digital Libraries are described to external parties, how appropriate resources are selected automatically, how text, image and audio (recorded speech) databases are searched, and how multimedia search results are displayed. Solutions deployed on a world-wide scale require a solid theoretical foundation capable of coping with significant heterogeneity, which we will develop. Every aspect of the proposed research will be tested, at the component level, and through a set of user studies covering research and resources produced by several sites around the world.

The result of the proposed research will be:

- ◆ a variety of methods and tool for metadata generation for different media;
- ◆ methods and tools for resource selection;
- ◆ methods and tools for merging ranked lists of items retrieved from the selected collections (data fusion);

All these methods and tools will be evaluated on different sources of data and cohorts of users.

3 Participant list

Partic. Role*	Partic. no.	Participant name	Participant short name	Country	Date enter project**	Date exit project**
C	1	University of Strathclyde	USG	UK	Start of project	End project
P	2	Universitaet Dortmund	UNIDO	D	Start of project	PM24
P	3	Universita' di Firenze	DSI	I	Start of project	End project
P	4	University of Sheffield	USFD	UK	Start of project	End project
P	5	Carnegie Mellon University	CMU	US	Start of project	End project
P	6	Universitaet Duisburg	UNIDU	D	PM25	End project

Note: the University of Duisburg replaces the University of Dortmund from PM25 to the end of the project. This is due to a relocation of the group involved in MIND. Since the UNIDU team is exactly the same as the UNIDO team, some sections of this document did not require any change (e.g. the innovation section). Changes and additions have been made to this document to indicate the effort of UNIDO and UNIDU in the different WPs and Deliverables.

4 Contributions to programme/key action objectives

There are many ongoing digitisation programmes in place across Europe in general, many of which are making available unique collections of textual, audio and visual material which would be otherwise inaccessible to the wider European community.

The proposed research directly addresses issues that arise when people have routine access to thousands of heterogeneous multimedia Digital Libraries distributed around the world. It will provide users with tool to enable a better resource selection and data fusion of data contained in several different multimedia Digital Libraries. It is therefore concerned with improving access to Europe's expanding repositories of cultural and scientific knowledge.

People need user interfaces that enable them to more fully exploit the multimedia information they find. The work carried out in MIND aims at improving the effectiveness of access to cultural and learning resources. It aims at demonstrating that it is possible to create high-performance federated Digital Libraries that integrate resources from many different sources and countries. It is therefore aimed at providing tools for federating content with navigation, search and retrieval functions and sharing internationally distributed resources and collections. Evaluation will be carried out using a variety of different cohorts of users and informative contents.

The project will also address and evaluate the need for standards to facilitate resource selection and fusion in terms of a common interface and metadata (resource descriptions), the results of which will be of interest to a number of players in the digital economy. The results of the project will be made available on a publicly available web-site as will prototypes, for testing and feedback from the wider community of information users and providers, also addressing the need for consensus on common specifications and practices for new data models, architectures, benchmarks and metrics, and test suites. An end-of-project workshop will enable dissemination of the results to a wider community and, in particular, to the industrial community. We hope that some of the technology developed in the context of MIND will be passed on to business/industrial exploitation, therefore improving the economic prospects of EU in the information technology market.

5 Innovation

Most of today's Digital Libraries are stand-alone information systems containing documents (or document metadata) in electronic form. As the number of Digital Libraries increases, dealing with numerous stand-alone Digital Libraries becomes unacceptable for the user. In order to satisfy an information need, s/he has to visit several Digital Libraries; searches have to be reformulated for each Digital Library; data about the same document is spread over different Digital Libraries such as bibliographic reference, citation and full-text databases. In order to overcome this situation, the concept of *federated Digital Libraries* has been proposed. Such a system should allow for searching across different Digital Libraries and merge the results for identical documents, thus giving the user the view of a single virtual digital library. However, different Digital Libraries do not only vary in the sets of documents and the kinds of metadata, they also use different schemas for representing this information and they provide different search functionality for accessing their content.

So far, two kinds of solutions have been proposed for searching in federated Digital Libraries:

- database-oriented approaches that apply the idea of federated databases; thus they are able to deal with the rich, and heterogeneous information structure of federated Digital Libraries, but fall short in handling the intrinsic vagueness and imprecision of information searches in Digital Libraries;
- information retrieval approaches that focus on vagueness and imprecision, but have neither addressed the heterogeneity problem nor considered other media than text.

In the proposed project, we want to start from information retrieval approaches and extend them for dealing with different kinds of media as well as handling heterogeneous Digital Libraries. We will develop methods for information search in federations of heterogeneous multimedia Digital Libraries, and we will evaluate effectiveness and efficiency of these methods. When a user submits a query to a federated Digital Library, first the system has to determine those Digital Libraries to which the query should be forwarded (*resource selection*). This step is based on some metadata about the content of each Digital Library (resource descriptions). Once the Digital Libraries have been selected and the query forwarded to them, the results returned have to be merged (*data fusion*) such that the overall retrieval quality is maximised.

Both resource selection and data fusion rely heavily on the availability of resource description describing the different Digital Libraries involved. We will develop methods for automatically deriving metadata for different media (text, facts, images, and speech). Whereas most methods proposed for solving this problem are based on the Digital Library content itself (thus assuming co-operative Digital Library systems with the appropriate functionality), we will also consider the case on non-co-operating Digital Library systems where only the query interface is available. For this case, appropriate methods for automatic acquisition and verification of resource descriptions will be developed.

For resource selection, all implemented methods so far are based on heuristic approaches and are restricted to either text or images as media. Here, we will start from a decision-theoretic model instead, that considers retrieval quality as well as a number of cost factors (response time, computing resources, and financial costs). Methods for applying this model will be devised, and will be applied to all kinds of media considered. The data fusion methods investigated in this project will be applicable to all kinds of media as well. Furthermore, we will investigate also different possibilities for visualising the merged results, where also non-linear rankings will be considered.

Today's Digital Libraries are rather heterogeneous with regards to schemas, retrieval capabilities and indexing methods, but most research on federated Digital Libraries so far has been restricted to the case of homogeneous Digital Libraries. Here we will deal with the different aspects of heterogeneity. Based on a new theoretical model that combines heterogeneity with uncertainty, appropriate methods for deriving metadata, for resource selection (including query translation) and for data fusion will be developed.

The methods developed will be evaluated on large multimedia Digital Library test-beds. Whereas most evaluations of this kind have been rather system-oriented, we will follow a user-centred approach, thus giving results that are more valid for the end-user.

6 Community added value and contributions to EU policies

The MIND project brings together partners at the European level who have access to the relevant digital resources which will be essential to prove that the proposed technological solutions will work across a variety of media types, held on multiple platforms and accessible through multiple national and international networks. The expertise and information resources of the European partners will be complemented by those of the American partner.

Although it is envisaged that MIND partners will not be involved in direct commercialisation the results of the project have the potential to facilitate the development of new business related to the creation, analysis, and distribution of information. Effective technologies for the identification of specific resources from a rapidly expanding and highly distributed collection of resources should improve the revenue generation possibilities for producers of digitised IPR and infomediaries. The project will also address and evaluate the need for standards to facilitate resource selection and fusion in terms of a common interface and metadata, the results of which will be of interest to a number of players in the digital economy.

MIND partners will also maintain a dialogue with social inclusion projects (e.g. the Glasgow Digital Library) learning resource projects (e.g. Clyde Virtual Library, FernUni Online, Nettuno Consortium, and NASA's Classroom of the Future) and related national digital library projects (e.g. eLib). It is therefore pursuing its research through contact and informal collaboration with other programmes in member states at an international level.

7 Contributions to community social objectives

As noted above, there are many ongoing digitisation programmes in place across Europe in general, many of which are making available increasing numbers of collections of textual, audio and visual material which would be otherwise inaccessible to the wider European and International community. Access to these diverse cultural and scientific materials is important for researchers, students and the citizen, as well as for commercial organisations. The potential to access these, often unique, resources, is beneficial in terms of the removal of geographical barriers, and the improvement in awareness of cultural heritage and diversity and exploitation of knowledge.

However, the exponential increase in the availability of these resources is counterbalanced by new problems for the end-user. Today, a person must know *where* to search, *how* to query different media, and how to *combine* information from diverse resources. All of these create new opportunity costs for the user, and, in extreme cases, information overload can reduce confidence in decision-making or lead to stress in the workplace (Information Fatigue Syndrome). As Digital Libraries continue to proliferate, in a variety of media, and from a variety of sources, these problems of *resource selection* and *data fusion* become major obstacles and sources of anxiety.

The proposed research will directly address some of these issues, investigating the design and development of techniques and tool for enabling more effective access to federated heterogeneous multimedia Digital Libraries distributed around the world. The aim is to enable users to exploit more fully the multimedia information available in these Digital Libraries. This will improve access to cultural and learning resources and hence contribute to the development of a better informed and skilled citizenry.

8 Economic development and S&T prospects

The exploitation of the work carried out in MIND will follow two directions: a research exploitation direction and a commercial exploitation direction.

Since the consortium consists of academic partners very active on in research, the research exploitation direction will concentrate on the integration of the results of the project with other current and future projects aimed at improving access to information resources. Partners are involved in a number of projects whose research objectives are related to those of MIND. Papers will be delivered at the key information retrieval and

Digital Libraries conferences and submitted to leading journals. We plan to publish results of the project at international Information Retrieval and Digital Library conferences (e.g. ACM-SIGIR, CIKM, ECDL, and ACM-DL) and journals (e.g. Information Processing and Management, Journal of the ASIS, ACM Transactions on Information Systems, Information Retrieval, and the International Journal of Digital Libraries). A workshop will also be held at the end of the project to demonstrate the results of the project to other academics.

The second approach will be pursued through promoting the availability of results to the information industry sector, rather than through direct commercialisation, nevertheless the commercialisation departments of the respective institutes will have access the work carried out in MIND. These departments are responsible for negotiating licensing and partnership agreements with commercial organisations and have a proven track of success in this task. Other activities aimed at promoting the availability of results to the information industry sector will include publishing overview articles in selected trade and practitioner magazines, talks given at corporate research laboratories, science parks, etc. The end-of-project workshop will also be aimed at potential end-users and potential business partners.

In order to highlight the availability of results the consortium will establish a web-site that will provide public access to selected deliverables including prototype tools, technical overviews and test results. This will ensure that researchers and potential business partners will benefit from the activities of the consortium as well as providing a channel for feedback on system functionality and effectiveness. The consortium coordinator (USG) will be responsible for setting up the web-site and for liaising with partners to populate the site with appropriate technical and promotional material. It is anticipated that the results of the project will be of interest to a number of businesses and users who are involved in the generation, communication and exploitation of information. These include standards bodies, portal and portal operators and other infomediaries, organisations involved in digitisation projects, and other research projects involved in resource discovery and identification.

In the following we report the specific directions of exploitation of the consortium partners.

University of Strathclyde (USG)

The MIND project represents an opportunity for the University of Strathclyde to develop additional skills and knowledge in the areas of multimedia, information retrieval and resource discovery. The University's Centre for Digital Library Research is currently involved in a collaborative project which has developed the Clyde Virtual University (CVU) involving four universities in the West of Scotland. The results of MIND should fit very well with the aims and objectives of CVU and would ensure wider exploitation by the academic community.

A second potential area of exploitation is integration with the results of the Glasgow Digital Library project (GDL). This project aims to establish the Glasgow Digital Library as a virtual co-library of the majority of public institutions in Glasgow. The long term aim is to create a wholly digital resource to support teaching, learning and research at all levels in the city, bringing together material currently separated by ownership and physical location. This will address both social inclusion and learning issues.

The University of Strathclyde has also set up an e-Systems Institute whose aim is to bring together expertise from the Science, Engineering and Business faculties in the areas of informatics. The Institute is building partnerships between the University and major ICT companies located in the West of Scotland. There will therefore be opportunities to demonstrate and publicise the results of MIND to the commercial community.

University of Duisburg (UNIDU)

The University of Duisburg will implement the methods for text and facts as part of the DAFFODIL system, currently being developed within its DAFFODIL project. This system will be open-ended for end users in order to access Digital Libraries of the computer science field, and it will undergo a thorough evaluation with real end users.

Università degli Studi di Firenze (DSI)

The University of Florence will explore a number of channels to promote dissemination of research results to the industrial community. The most promising of these channels is the Master in Multimedia that is organised

by the DSI of the University of Florence, in cooperation with RAI - Radio Televisione Italiana (public television company) and the Multimedia Library of Tuscany. The Master in Multimedia is a one-year post-graduate course of advanced teaching and experimentation, including a final internship within a company, a firm, or an institute. In this context, companies participating to the Master in Multimedia are interested in the transfer of research results in industrial applications. DSI will promote dissemination and communication of project results to industrial subject involved in the Mater in Multimedia.

DSI will also contribute to project dissemination by hosting and organising the workshop that will be held at the end of the project. The more central location in Europe of Firenze and its closeness to a number of less-favored regions will facilitate the dissemination of the results of the project where they are most needed.

University of Sheffield (USFD)

Sheffield University has a university company scheme for the protection and exploitation of IPR in special cases. The main method of exploitation is dissemination of expertise and research results, to business by consultancy and availability for contractual research and development, and otherwise by the normal dissemination channels of publication, conference papers, organised workshops and courses, and the world wide web. Sheffield's Information Studies Department and ILASH have a substantial track record in all these modes of dissemination and university-style exploitation. We shall participate vigorously in the protection of IPR from this project by copyright and patent, and will participate fully in the final Technical Implementation Plan (TIP).

Carnegie Mellon University (CMU)

Carnegie Mellon University is working with a number of government partners (U.S. Geological Survey, U.S. Department of Commerce, General Services Administration/Regulatory Information Service Center, and the U.S. Library of Congress). CMU will produce a prototype of a system for accessing distributed, heterogeneous, government information, and demonstrate its utility. Building this system will be an important part of evaluating the research, and the first step in transferring the new technology to government systems.

The four government agencies will contribute databases unique to their agencies, including both text and tabular data. The agencies will also help to evaluate the prototype demonstration system with respect to the specific information needs of their agencies. The USGS will be the lead government agency, because of its expertise in interoperability standards such as Z39.50 / GILS and the Advanced Search Facility, and will contribute to enabling the prototype demonstration system to interoperate with information systems from each of the four government agencies. The USGS will provide examples of cross-organizational information needs within the government, and examples of "citizen queries" that are not addressed adequately by current systems.

9 Workplan

9.1 General description

The research in MIND addresses issues that arise when people have remote access to thousands of heterogeneous and distributed multimedia Digital Libraries. In order to cope effectively with such masses of knowledge the first task to solve is *resource selection*. Among the Digital Libraries accessible those have to be selected which most probably contain relevant documents with respect to the users information need. The second task is the distributed retrieval task. The user's information need has to be processed against the resources selected in the first step. The third task is to fuse the results obtained from the selected resources. The retrieval results need to be merged into a single result and visualised in a proper way for the user.

MIND will develop models and algorithms for solving these tasks and implement them in order to provide a tool set as depicted in the generic architecture in figure 1.

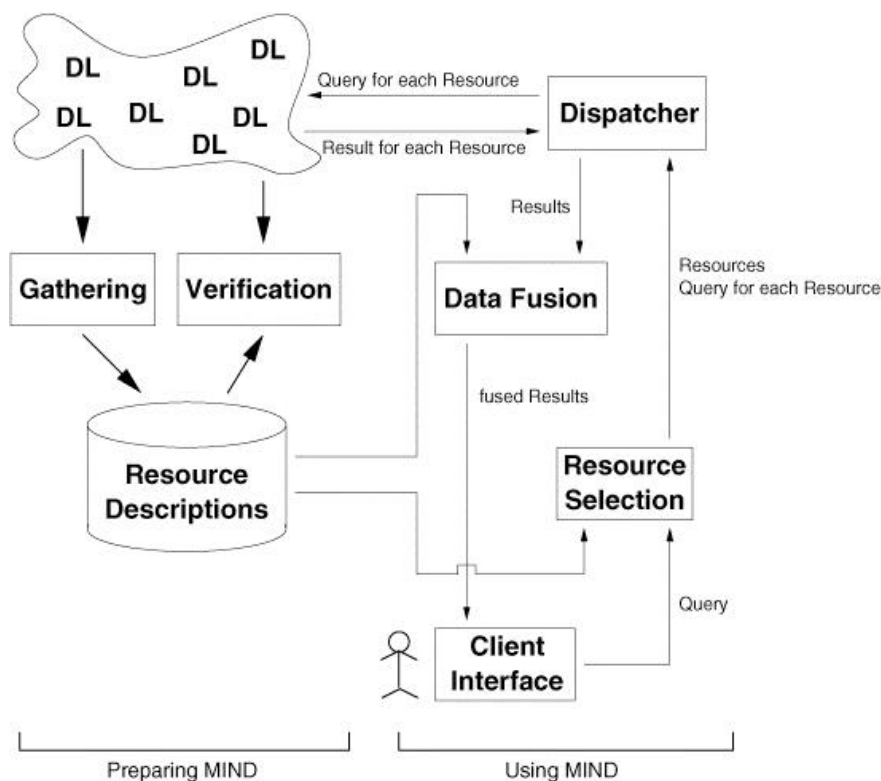


Figure 1: generic architecture of the MIND prototype.

Following the generic architecture the project work is partitioned into four phases:

1. Specification (PM 1-6)
2. Resource selection and data fusion from homogeneous Digital Libraries (PM 7-16)
3. Heterogeneity (PM 17-24)
4. Evaluation (PM 18-30)

Following is a description of these phases and an overview of the milestones of the project.

1. Specification

In the specification phase, first a basic architecture for resource selection and data fusion will be defined. Also, the type of Digital Libraries to be considered throughout this project will be specified, and possible application areas identified. Since both resource selection and data fusion require appropriate resource descriptions, the format of these descriptions, i.e. the kind of metadata resource descriptions consist of, has to be specified. For this purpose, the structure of resource descriptions for the three different media types considered in this project (text, images, and speech) will be defined. At the end of this phase, a resource description specification will be available. The major active workpackages in this phase are WP1 (Architecture) and WP2 (Content Metadata). WP2 will continue after this phase, dealing with specific aspects of resource descriptions: their automatic acquisition and automatic verification as well as with the problem of resource descriptions at different granularities.

2. Resource selection and data fusion from homogeneous Digital Libraries

The second phase focuses on resource selection and data fusion in the case of homogeneous Digital Libraries. For resource selection, first a theoretical framework will be developed. Based on the decision-theoretic model from UNIDU, different strategies for estimating the three classes of parameters have to be devised. Subsequently, these strategies will be implemented for the different media. For data fusion, different types of information will be considered namely query-specific data, document-specific data and collection-specific data; for each of these cases, appropriate fusion strategies will be developed. For both resource

selection and data fusion, the quality of the strategies developed will be evaluated using small test-beds. At the end of this phase, tools for resource selection and data fusion for homogeneous Digital Libraries will be available. The major active WPs in this phase are WP2 (Content Metadata), WP3 (Resource selection) and part of WP4 (Data fusion). The continuation of WP4 after this phase will deal with the visualisation of fused results. Appropriate tools will be ready for the evaluation phase.

3. Heterogeneity

Phase 3 addresses the issue of heterogeneity, i.e. when an application area involves Digital Libraries that use different media, different indexing methods for the same media or different database schemas. For this case, additional mappings are required that allow for transformations between the different representations; furthermore, the resource selection and data fusion strategies developed for the homogeneous case have to be modified appropriately. The different strategies will be evaluated using small test-beds. At the end of this phase, tools for resource selection and data fusion for heterogeneous Digital Libraries will be available. The major active WP in this phase is WP5 (Heterogeneity).

4. Evaluation

The final phase deals with evaluation, WP6. For this purpose, first an appropriate test-bed will have to be selected, and the necessary resource descriptions for the test-bed will have to be collected. Parallel to this activity, the tools developed in the previous phases that are applicable on this test-bed will have to be integrated in the form of a prototype system. After completion of these activities, a user-centred evaluation of the prototype system will be performed. The major outcome of this phase will be the evaluation report describing the evaluation results.

Key Milestones

The key milestones of the project are months 6, 12, 16, 24 and 30 (end of project) and their associated deliverables. For the four phases of the project these are:

1. Phase 1: The specification of the architecture of the MIND prototype and the identification of target categories of users and Digital Libraries' informative contents (D1.1, D1.2); the specification of resource description formats for text, image, and speech resources (D2.1).
2. Phase 2: The specification of models and methods for resource selection and data fusion for homogeneous resources (D3.1, D4.1).
3. Phase 2: The implementation of tools for automatic extraction and verification of resource descriptions (D2.2); the implementation of tools for resource selection and data fusion for homogeneous resources (D3.2, D4.2).
4. Phase 3: The definition of methods for automatic generation of surrogates and hierarchical overview of fused results, accompanied by prototype tools (D4.3); The specification of methods for resource selection and data fusion for heterogeneous collections, accompanied by prototype tools (D5.1, D5.2).
5. Phase 4: A prototype system for resource selection and data fusion, which will have been subjected to an extensive evaluation by various user groups (D6.1, D6.2, D6.3), the main findings of which will be publicised and discussed at an end of project workshop (D7.4).

In addition to the final evaluation, also limited evaluations using recall and precision measurements will be performed in WPs 3, 4 and 5. The final deliverables for each of these WPs will contain the corresponding evaluation results.

9.2 Workpackage list

Work-package	Workpackage title	Lead partner	Person-months	Start month	End month	Deliverable No
WP1	Architecture, content and context	UNIDO	10 (+ 0.5)	1	3	D1.1 - D1.3
WP2	Content metadata	DSI	22	4	16	D2.1 - D2.2
WP3	Resource selection	UNIDO	20 (+ 1)	4	16	D3.1 - D3.2
WP4	Data fusion	USFD	32 (+ 1)	4	21	D4.1 - D4.3
WP5	Heterogeneity	USFD	38.7 (+ 3.5)	7	24	D5.1 - D5.2
WP6	Evaluation	USG	39.3 (+ 4.5)	16	30	D6.1 - D6.3
WP7	Dissemination	USG	2 (+ 3.75)	1	30	D7.1 - D7.4
WP8	Exploitation	USG	2 (+ 3.75)	1	30	D8.1 - D8.3
WP9	Project Management	USG	20 (+ 10.5)	1	30	D9.1 - D9.7
	TOTAL		186 (+ 28.5)			

9.3 Workpackage descriptions

WP1: Architecture, Content and Context (UNIDO)

Workpackage number :	WP1	Start date or starting event:					PM1
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	2 (+ 0.5)	2	2	2	2	0	

Objectives

This WP aims at the definition of: a framework architecture for resource selection and data fusion that forms the basis for this project. Furthermore; the types of data sources considered as well as the user groups targeted at will be defined. In addition a development and evaluation plan will be drawn which will assure coherence of development and proper involvement of users at the different stages of the project evaluation.

Description of work

The major focus of the project is the development of methods and tools for resource selection and data fusion. However, in order to allow for the integration of these tools in concrete settings (especially for the purpose of evaluation within the project), it is necessary to agree upon the basic architecture as well as the data sources and user groups to be considered. This set of activities will be carried out in this WP.

Risks

The MIND project will develop methods for resource selection and data fusion in multimedia Digital Libraries. Whereas text-only Digital Libraries are widely available today (comprising also a limited amount of images, as they usually occur in text documents), libraries with a large share of non-textual content can be found in very limited domains only. On the other hand, the collections of multimedia files on the Web made accessible by popular internet search engines can not be viewed as multimedia Digital Libraries, due to their heterogeneous nature and the lack of any metadata. Given this situation, the MIND project may have problems in the acquisition of appropriate multimedia test data and relative user group for development and evaluation.

In order to address these issues, we will take appropriate steps at an early phase of the project. For the multimedia collections, we will set up our own testbeds locally. We have already identified a number of available DLs and relative user groups that we could use. In particular, for DLs:

1. paintings database: about 400 color images of XX century paintings;
2. photographs: about 100 grayscale images of generic subjects;
3. advertising video clips: about 100 taken from Italian TV;
4. news video clips: about 200 taken from Italian TV;
5. sports video clips: under construction for the EC ASSAVID project, mostly from BBC sports;
6. BBC Word Service data: speech from radio;
7. TREC text collections: CDs 1,2,3; newswire, magazine, and government documents; 500.000-1,250,000 documents, about 100 databases (also 20 Gb TREC VLC corpus; Web data; 5,000,000 documents);
8. TREC speech data: several Gigabytes of recorded speech (in addition USG is involved in the new video track of TREC, to start in 2001).

For each of these collections, there are also specific user groups. We have already preliminary agreements with some user groups:

1. ALINARI photographic archives in Florence;
2. RAI Italian TV: news video, sports video;
3. BBC UK: sports videos;
4. KATAWEB La Repubblica: the most important news on-line in Italy.

Thus, even if appropriate multimedia Digital Libraries are not publicly available, we will perform evaluations with the local testbeds, but with real users.

Task 1.1: System architecture (UNIDO). The purpose of this task is the definition of a framework architecture. Starting from the generic architecture shown in Section 9.1, the design of the system architecture will be refined to the extent that the different tools developed throughout the project can be integrated. For this purpose, the interfaces between the different modules have to be specified at a media-independent level whereas the media-dependent formats will be defined later (in WPs 2 and 5). In addition an agreement in the consortium will need to be reached regarding the platform of development, the programming languages to be used, the communication

protocol, etc. This agreement will be part of the development and evaluation plan (see below).
Task 1.2: Content and user groups (USG). With respect to content and user groups, Digital Libraries may vary to a great extent. In this task, the types of content to be considered will be specified, and typical representative DLs of these types will be identified. In order to be considered by the test-beds of this project, application areas where several multiple Digital Libraries are available will have to be identified. Furthermore, potential user groups for these federations of Digital Libraries will be identified as well as concrete user groups that will be engaged for evaluating the testbed. A development and evaluation plan will be drawn that will detail the involvement of users in the evaluation. In particular, this plan will address questions related to the scope/range of involvement of user groups at the different stages of the project, and the provision of utilising user feedback effectively and systematically. Questions related to where, when and how often resource descriptions are gathered and verified will also be addressed.

Deliverables

- D1.1: Test-bed architecture specification.
- D1.2: Identification of content and user groups.
- D1.3: Development and evaluation plan.

Milestones and expected result

PM3: D1.1, D1.2 and D1.3 completed.

WP2: Resource Descriptions (DSI)

Workpackage number:	WP2	Start date or starting event:					PM4
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	3	3	3	2	11	0	

Objectives

The aim of this workpackage is to develop a comprehensive set of methods and prototype tools for the definition and extraction of resource descriptions for representation of text, audio and image content. Criteria for characterising resources have to be specified. A major goal of this workpackage will also be to assemble representative text, audio and image material that can be used for system development and testing.

Description of work

For resource selection and data fusion knowledge about the actual content of a given resource must be available. The form of this metadata is media-dependent and may also vary in its granularity. In this WP first a set of criteria for assembling such metadata to resource descriptions has to be defined. Criteria include for example storage and update cost, precision of description, and support of search predicates. The major task of this WP is the development of a variety of resource description generation methods for different media and with different granularity that perform well with respect to the criteria. Test materials will be collected from available archives so as to cover a wide and significant range of cases in the application context. Representative examples of this material will be compiled and disseminated to the project partners.

Risks

The risk of non-availability of sufficient quantities of resource descriptions and metadata to be included in the resource descriptions will be assessed. Previous experience by some of the project partners shows that this risk is low, but in case it becomes real, alternative ways of automatically deriving Digital Library resource descriptions will be explored.

Task 2.1: Text Metadata: Automatic Acquisition of Resource Descriptions (USG). Controlled vocabulary terms and information provided by Digital Libraries will be used to define *resource descriptions* for text collections. *Query based sampling* will be adopted for acquiring accurate resource descriptions from resources that don't cooperate. This will support the extraction of effective resource description which is mandatory in order to enable an accurate resource selection (WP3).

Task 2.2: Text Metadata: Automatic Verification of Resource Descriptions (CMU). Query based sampling will also be used to enable verification of resource descriptors provided by resources that cooperate. Verification will be carried out either by checking the frequency of resource descriptors in samples of retrieved documents or by extracting resource descriptors from the digital library (as in Task 2.1) and comparing the two sets of descriptors.

Task 2.3: Audio Metadata: Automatic Acquisition of Resource Descriptions (USFD). A complete theory based on *non-Bayesian reasoning* will be formalised to describe uncertainty for audio metadata. This theory will

enable combination of the different sources of uncertainty that are associated with audio metadata generation (mainly related to word recognition errors, and confidence on the source of data). The use of query based sampling for the extraction of resource descriptors from audio archives relies on a speech recognition engine that verifies the presence of the query words in retrieved documents. The effect of the query types on the estimated quality of the collection will be investigated.

Task 2.4: Image Metadata: Automatic Acquisition of Resource Descriptions (DSI). Methods will be defined for the extraction of colour, shape and texture content descriptors form image data. Content descriptors of individual items will be subject to a clustering process so as to extract descriptors of image categories. The clustering process can be carried out through the organization of content descriptors into a metric access index. The different nodes of the tree structure can be used to derive a descriptor of the content of all the items represented in its sub-tree. Hence, descriptors included in the highest nodes can be used to derive resource descriptors according to colour, shape and texture content.

Task 2.5: Metadata at Different Granularities (UNIDO). Specific criteria will be developed to derive metadata for facts and metadata at different granularities. The effectiveness of using condensed metadata will be investigated with regard to the kind of search predicates that can be supported.

Deliverables

D2.1: Definition of content metadata structure for text, audio and images.

D2.2: Prototype tools for content metadata extraction and verification at different granularities.

Milestones and expected result

PM6: D2.1 completed.

PM16: D2.2 completed.

WP3: Resource Selection (UNIDO)

Workpackage number :	WP3	Start date or starting event:					PM4
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	3	10 (+ 1)	2	2	3	0	

Objectives

Based on the decision-theoretic model developed by UNIDO, strategies for estimating the parameters have to be developed. The Parameter estimation procedures are applied to resource descriptions for text, image, and speech collections. Tools that implement resource selection for these media are to be implemented accordingly.

Description of work

Resource selection deals with the problem of identifying the best databases for answering a query. For this purpose, the decision-theoretic model developed by UNIDO will be used. For each database that can be accessed, this model considers the expected retrieval quality, the expected number of relevant documents and cost factors for query processing and document delivery. Given these parameters, a divide-and-conquer algorithm computes an optimum selection of databases. The major task here is the estimation of the parameters of the model. Whereas estimation of cost parameters may be fairly easy, the crucial problem is the estimation of the expected retrieval quality of the different databases. For this purpose, the suitability of the different forms of metadata developed in WP2 will be investigated. In order to simplify the problem of parameter estimation, we will first consider specific cases where not all parameters are required, and later work towards the full model.

Task 3.1: Resource Selection Framework (UNIDO). Based on the decision-theoretic model developed by UNIDO, different strategies for estimating the three classes of parameters have to be developed. The expected retrieval quality can be estimated by considering the expected precision of the retrieval system for the different search predicates used in the query formulation. The number of relevant documents in a database can be estimated based on the principle of retrieval as uncertain inference, where the retrieval process is subdivided into an uncertain inference part and a relevance estimation part. For the first part, the similarity between query and documents computed by different media-specific retrieval methods has to be considered. For the second part, media-independent methods for estimating the probability of relevance have to be investigated, e.g. by using relevance feedback.

Task 3.2: Similarity of a Query to Each Text Resource (USG). For text, CMU has developed a resource selection method that compares favourably to competing approaches. This method will be further improved such that its results become less biased by database size. Also, it will be investigated how this approach scales with the number of databases. The current selection method relies on probabilistic and vector retrieval methods, whereas language model approaches have become very popular in text retrieval recently; thus, the combination of this type of models with selection will be investigated. Preliminary research results show that results of distributed retrieval can be improved by reorganising databases such that their content becomes more homogeneous. Our research interest is to apply the lessons learned from reorganising databases to situations where it is not possible to reorganise the data. Another area of research will deal with the exploitation of query expansion methods for resource selection; here the problem is to choose the appropriate resource(s) for the expansion steps.

Task 3.3: Similarity of a Query to Each Audio Resource (USFD). Although text and speech retrieval use very similar methods, there are two major differences that have to be taken into account for resource selection. 1) Speech recognition systems have a limit to the size of vocabulary they can recognise. We will examine if out of vocabulary words present a problem for collection selection, and if they do, will explore ways of adjusting query expansion techniques to alleviate the problem. 2) Selecting collections from a heterogeneous set with varying recognition rates presents some problems, since current selection techniques assume equal recognition quality. Using the estimations of audio collection error generated in WP2, we will adjust collection selection techniques to allow for this error.

Task 3.4: Similarity of a Query to Each Image Resource (DSI). Visual retrieval systems support retrieval by visual content by directly addressing image visual features such as colours, shapes, textures and spatial relationships. These features all address the syntactic level of images, whereas text and speech retrieval methods consider words as semantic entities. On the other hand, text and speech retrieval only consider the presence or absence of features (words), whereas image retrieval is based on similarity of feature values. As a first step, the application of methods developed for text onto the image domain will be investigated. In a second step, the problem of feature value similarity will be addressed.

Deliverables

D3.1: Specification of the resource selection framework.

D3.2: Prototype tools for resource selection.

Milestones and expected result

PM12: D3.1 completed.

PM16: D3.2 completed.

WP4: Data Fusion (USFD)

Workpackage number :	WP4	Start date or starting event:					PM5
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	10 (+ 1)	5	10	4	3	0	

Objectives

We will investigate two aspects of data fusion: how to fuse multimedia results returned from multiple Digital Libraries into a single result, and how to display results visually. Depending on the information available, different fusion strategies have to be developed. After fusing the retrieval results, we will investigate possibilities for their display. The documents retrieved by the system could be conventional text documents, text-free documents (e.g., images), and audio segments processed into text by a speech recogniser.

Description of work

Once the databases for a query have been selected and the queries have been sent to them, data fusion deals with the problem of forming the overall result. In the standard case, this will be a single ranked list, where the corresponding retrieval quality depends heavily on the strategy used for combining the rankings from the different databases. For this purpose, data from different sources will be used to influence the combination, namely query, document and collection-specific information from the databases. Depending on the information available, different fusion strategies have to be developed. Having generated a single retrieval result from multiple sources the result needs to be prepared for being displayed to the user.

Task 4.1: Use of query specific information for data fusion (DSI). UNIDO will contribute a means of handling data fusion with respect to query conditions relating to facts. DSI will concentrate on the development of criteria and methods to derive a compound matching score accounting for different visual features, such as colour, texture and shape, as well as textual and aural features. In this context, the analysis will focus on the use of effective retrieval interfaces to allow users to specify the relevance of the different elements used in the query, so as to drive the fusion strategy toward a result which best matches users' expectations.

Task 4.2: Use of document specific information for data fusion (USG). USG will tackle the problem of data fusion by providing its expertise in generating and using document associations to enhance retrieval of search results. Document associations can be pre-existent, like for example document citations, or can be built on-the-fly by discovering similarities in document characteristics, via term distributions or through user preferences. Associations can also be divided into factual and semantic categories. Factual associations reflect assigned characteristics of documents, like authorship, publishing date, author affiliation, or document citations. Semantic associations, instead, reflect content similarity between documents. We believe that the correct exploitation of these types of associations between documents can enhance the effectiveness of Digital Libraries.

Task 4.3: Use of collection specific information for data fusion (USFD). USFD will address the issue of using collection specific information to drive fusion strategies. Currently, most data fusion research has made the assumption that the word characteristics of each collection being retrieved from, are the same. When retrieving from heterogeneous collections, this assumption has been shown to be false. We propose to examine this issue by exploring collection analysis methods and alternative document ranking schemes that will first, identify and second, account for the effect described here.

Task 4.4: Automatic generation of surrogates for the display of fused results (USG). After fusing the retrieval results, USFD and USG will investigate possibilities for their display. The documents retrieved by this system could be a combination of conventional text documents, text-free documents, and audio segments processed into text by a speech recogniser. Existing means of presenting retrieved documents is as a list of text surrogates (e.g. document title, passages extracted from body text, etc) ranked in order of relevance. Most documents being retrieved will contain or be described by text in some form. However, for non-text documents, text surrogates will not be available. In this task we will investigate a method for automatically generating a readable surrogate for these document types (USFD and USG).

Task 4.5: Hierarchical overview of the fused results (USFD). As an alternative to presenting a list of document surrogates as the result of a retrieval run, USFD and DSI will build a system that will present a thesaural like hierarchical organisation of words and phrases extracted from the retrieved documents. The hierarchy attempts to reflect the topical structure of the documents. A system to build this structure is already in development. In this project efforts will concentrate on applying the structure to the documents of the MIND collection.

Deliverables

D4.1: Definition of methods for data fusion.

D4.2: Prototype tools for data fusion.

D4.3: Methods and tool for automatic generation of surrogates and hierarchical overview of fused results.

Milestones and expected result

PM12: D4.1 completed.

PM16: D4.2 completed.

PM21: D4.3 completed.

WP5: Heterogeneity (USFD)

Workpackage number :	WP5	Start date or starting event:					PM8
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	4(+ 0.5)	10.1(+ 1)	6.2	12.4 (+ 2)	3	3	

Objectives

The major task of this workpackage is the development of mappings coping with the heterogeneity of the Digital Libraries. Cross media mappings will be developed which allow for querying in one media and getting results in different media. For a single media mappings between different representations will be developed. The tools for resource selection and data fusion as developed in WP3 and WP4 are then extended accordingly, in order to be able to cope with heterogeneity.

Description of work

While in the first phase of the project, we assume a uniform structure of databases, i.e. all databases have the same schema, apply the same indexing methods and provide the same resource description and result information. In the second phase effects due to heterogeneity will be investigated. The major task of this workpackage is the development of the mappings required for coping with heterogeneity. Cross-media mappings will allow for querying in one media and getting results in a different media. For a single media, mappings between different representations will be developed. In the most general case there are different schemas which have to be transformed into each other.

Task 5.1: Cross-media mappings (USFD). USFD will tackle the issue of deriving a means of mapping the indexes (i.e. schemas) of one media collection into another. The areas to be addressed are mapping speech and images into a textual form.

Task 5.2: Heterogeneity of representation (DSI). For all the media considered, there is the problem that Digital Libraries differ in the media representations that are derived from the indexing process. Thus, both the type of resource descriptions as well as the implementations of the query predicates may differ. Resource descriptions can be obtained automatically, using query-based sampling, or be provided by a Digital Library that follows a cooperative protocol. The cooperative solution suffers from the problem that it is not possible to directly compare occurrence statistics produced by different resources (resource-specific descriptions). Heterogeneity is not a problem for resource descriptions acquired with query-based sampling. The resource selection service can use the same indexing algorithm on documents from different resources, which results in statistics for different databases that are directly comparable. Query-based sampling enables a resource selection service to determine for itself the relative frequency of terms in different resources. It is an open research question whether a hybrid approach, based on resource descriptions provided cooperatively and metadata acquired by query-based sampling, is superior to either alone.

Task 5.3: Heterogeneous schemas (UNIDU). In order to deal with heterogeneous schemas of different Digital Libraries, UNIDU will use a logic-based approach. For representing the schemas, the Resource Description Framework (RDF, proposed by the W3C) will be employed, since it is anticipated that it will be used by future Digital Libraries to describe their schemas. The RDF descriptions then will be mapped into an object-oriented logic, for which appropriate transformation algorithms will have to be developed. As a representation language, we will extend the probabilistic object oriented logic POOL such that it supports typing and inheritance. Based on the logical representation, mappings between different schemas can be generated automatically. Strategies for computing imprecise mappings between different schemas and for considering vague predicates have to be developed.

Deliverables

D5.1: Definition of methods for resource selection and data fusion for heterogeneous collections.

D5.2: Prototype tools for resource description extraction, resource selection and data fusion for heterogeneous collections.

Milestones and expected result

PM28: D5.1 and D5.2 completed.

WP6: Evaluation (USG)

Workpackage number :	WP6	Start date or starting event:					PM22
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	8 (+ 2.5)	0	7	7 (+ 1)	7	10.3 (+1)	

Objectives

In addition to individual groups evaluating their components, there will be a large overall evaluation, coordinated by USG. For this purpose the tools developed in Workpackages 2 to 5 have to be integrated in the MIND prototype system. A test-bed of multimedia Digital Libraries will be build that will mimic in all aspects but the size a real distributed multimedia digital library. The results of the user-centred evaluation will be published on the end-of-project workshop.

Description of work

Having developed and evaluated the components of the generic system architecture they need to be integrated into the MIND system prototype. An overall system evaluation is then needed to show how the developments of the MIND project help users in accessing distributed, heterogeneous multimedia Digital Libraries. The evaluation will draw on USG’s expertise in the field of user-centred evaluation and will take place within a large realistic setting involving a number of heterogeneous multimedia Digital Libraries. This will draw on elements of the methodology developed by USG in the context of its participation to the ESPRIT Working Group “Mira” on the evaluation of interactive multimedia information retrieval applications.

The final evaluation will be carried out in collaboration with the user groups identified in WP1 according to the development and evaluation plan, where their involvement will have been detailed.

Risks

Concerning the scalability issue, we will have no problems in the text area. CMU currently is building a new test collection with about 65 GB of data, to be subdivided into several thousand databases. However, for the image and speech collections, even when the datasets are available, we may not be able to process these large quantities of data. If this problem occurs, we will perform a partial evaluation for the case of images, by performing a scalability test only for those features that can be efficiently computed. For speech, we will exploit the similarity between text and speech retrieval: we will test the deviation between the scalability of text and speech methods on a medium size testbed, thus giving us an empirical basis for extrapolation towards large collections. Thus, the project will test the scalability of the methods for text.

With regards to images and video, one must consider that in content based retrieval from image databases focussing on test database size can be misleading. While size is mandatory for assessing scalability of retrieval performance (and therefore to test effectiveness of indexing), it is not for similarity matching (in this case we have to check the similarity assessment by the user against the one performed by the system). We therefore need a limited set of images, appropriately selected, so that the user for each test query can define the k-nearest images and their similarity rankings, which is almost impossible if the database has a large size. For content based retrieval based on intermediate-higher level features (meaningful combinations of low-level features, or object/context classification) size is relevant just for taking into account the variability of image data. This does not necessarily leads to huge archives but rather to an accurate selection of them.

Task 6.1: Integration (UNIDU). The tools for resource selection and data fusion will be integrated into the MIND system prototype. In a preliminary integration tools which cope with homogeneous resources only will be used. Later in the project, when the components for dealing with heterogeneous resources become available, they will replace the prior ones.

Task 6.2: Gathering of data (DSI). DSI will construct and provide colour, texture and shape test image collections. USFD will provide audio and speech recognised collections. CMU will provide a selection of text collections and build a set of tasks, comprising information needs, queries, interaction models, and a framework for the evaluation of the user satisfaction of such tasks. USG will provide the general framework for the task evaluation. UNIDU and DSI will be involved in defining the evaluation criteria: DSI looking at the construction and definition of suitable measures of retrieval efficiency and effectiveness in image retrieval; and UNIDU examining classical (batch-oriented) retrieval measures as well as modern user-oriented measures for consideration.

Task 6.3: User-centred evaluation of the prototype systems (USG). Once the testing environment is in place and some suitable measures of performance are identified, USG will coordinate the user-centred evaluation of the prototype systems developed. USG will also co-ordinate the analysis of the data gathered during the evaluation.

Statistical analysis techniques will be used to isolate the different variables involved in the evaluation and to test the effectiveness hypotheses formulated. Such analysis will provide useful suggestions on how to improve the architecture, the interfaces and the user-interaction styles of the prototype systems.

Deliverables

D6.1: Components evaluation and integration.

D6.2: Testbed for resource section and data fusion for multimedia Digital Libraries.

D6.3: User-centred evaluation of the MIND system

Milestones and expected result

PM33: D6.1, D6.2, and D6.3 completed.

WP7: Dissemination (USG)

Workpackage number :	WP7	Start date or starting event:				PM1
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU
Person-months per participant:	(1.25)	(0.75)	1.25	(1.25)	.75	(0.50)

Objectives

To ensure widespread awareness of the results of the MIND project in order to create opportunities for commercialisation of these results and to encourage integration of the results with other current and future projects aimed at improving access to information resources.

Description of work

In order to highlight the availability of results the consortium will establish a web-site which will provide public access to selected deliverables including pilot retrieval systems, technical overviews, test results and promotional material. This will ensure that other researchers will benefit from the activities of the Consortium as well as providing a channel for feedback on system functionality and effectiveness.

This will be the principle route for involvement of, and evaluation by, users. The consortium partners have close links with the key European and North American Information Retrieval Special Interest Groups and will use these links to promote the results of MIND and encourage end-user evaluation. It is anticipated that the IR community will be a good source of constructive criticism and evaluation of the MIND tools. Papers will also be delivered at the key information retrieval conferences and submitted to leading journals. Overview articles will be submitted to selected trade and practitioner magazines. We plan to publish results of the project at international Information Retrieval and Digital Library conferences (e.g. ACM-SIGIR, CIKM, ECDL, and ACM-DL) and journals (e.g. Information Processing and Management, Journal of the ASIS, ACM Transactions on Information Systems, Information Retrieval, and the International Journal of Digital Libraries). Contact will also be made as appropriate with relevant standards bodies, organisations involved in digitisation projects, and other research projects involved in resource discovery and identification.

A workshop will also be held at the end of the project to demonstrate the results of the project to academics, potential end-users and to potential business partners. In addition, in the second year, another workshop will be organised in the context of a major international conference held in USA (the most appropriate venue will be chosen at the appropriate time). This workshop will be related to the general area of research of resource selection and data fusion and will present the results of MIND in relation to similar work being carried out in USA and Europe. The Consortium coordinator (USG) will be responsible for setting up the web-site and for liaising with partners to populate the site with appropriate technical and promotional material. It is anticipated that the results of the project will be of interest to a number of businesses and users who are involved in the generation, communication and exploitation of information. These include standards bodies, portal and vortal operators and other infomediaries, organisations involved in digitisation projects, and other research projects involved in resource discovery and identification. The coordinator will take responsibility for drawing a dissemination and use plan at PM6, which will include a clear commitment to investigate business plans and concrete integration strategies with other EU and national project (see section 10 on clustering).

Deliverables

D7.1: Development of web-site to support project promotion. Selected deliverables from other tasks will be made available on this web-site, as will copies of papers submitted to journals and conferences.

D7.2: Project presentation.

D7.3: Dissemination and use plan.

D7.4: Workshop to disseminate and discuss project findings.

Milestones and expected result

PM3: D7.1 and D7.2 completed.

PM6: D7.3 completed.

PM36: D7.4 held in Florence, hosted by DSI.

WP8: Exploitation (USG)

Workpackage number :	WP8	Start date or starting event:					PM1
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	(1.25)	(0.75)	1.25	(1.25)	.75	(0.50)	

Objectives

To identify opportunities for commercialisation of the results and to encourage integration of the results with other current and future projects aimed at improving access to information resources.

Description of work

The consortium main method of exploitation will be through dissemination of expertise and research results, to business by consultancy and availability for contractual research and development, and otherwise by the normal dissemination channels of publication, conference papers, organised workshop, and the world wide web (see WP8). Exploitation will therefore concentrate on the integration of the results of the project with other current and future projects aimed at improving access to information resources, and on promoting the availability of results to the information industry sector and end-users, rather than direct commercialisation. Potential projects include Clyde Virtual University (CVU), Glasgow Digital Library (GDL) and DAFFODIL. In addition our American partner will be seeking to transfer technologies to the its government partners. The first approach is intended to address social inclusion and learning opportunities issues through facilitating access to relevant information resources and developing a more informed citizen, consumer or employee. The second approach will be effected through the exploitation and commercialisation departments of the respective institutes. These departments are responsible for negotiating licensing and partnership agreements with commercial organisations. Other promotional channels used will be portals such as the Scottish Research Information System and the UK Higher Education Mall which are to act as showcases for intellectual property generated by HEIs.

The Consortium co-ordinator will be responsible for setting up a web-site and for liaising with partners to populate the site with appropriate technical and promotional material. The coordinator will also be responsible for the production of a promotional brochure to encourage exploitation of MIND results.

A technology implementation plan (TIP) will evolve through the lifetime of the project and will highlight innovative features of the developed technologies, market opportunities and, where feasible, comparisons of functionality and performance. In addition, a marketing prospectus, prepared by the coordinator, will draw the business exploitation plan reflecting more specifically the needs of the project.

Deliverables

D8.1: Technology implementation plan (TIP).

D8.2: Marketing prospectus highlighting key outputs of MIND.

D8.3: Exploitation and business plan.

Milestones and expected result

PM36: D8.1 Final TIP written. D8.3 developed.

PM30: D8.2 produced

WP9: Project management (USG)

Workpackage number :	WP7	Start date or starting event:					PM1
Participant number:	USG	UNIDO	DSI	USFD	CMU	UNIDU	
Person-months per participant:	15 (+ 5.5)	(1.75)	2.5	(2.5)	2.5	(0.75)	

Objectives

To undertake the administration of the project and to provide the necessary communication and control mechanisms between the partners to ensure their required co-operation to fulfil the project objectives to the desired quality.

Description of work

This WP will run throughout the duration of the project. It will ensure that the various partners are in regular communication and that they can share results and deliverables inside the consortium, in order to fulfil its declared objectives. Regular Project Co-ordination meetings will be held to exchange information between the partners and to share their experiences. The coordinator will also ensure that deliverables and periodic reports are delivered in time, that progress is in line with declared milestones, and that corrective action is taken in the event of unforeseen problems which may affect the successful completion of any task or work package.

Project management carried out by USG will also tackle, in concert with the other partners, any issue of significant risk in the different stages of the project and will draw contingency plans. Significant risks could be related to the being able to demonstrate the scalability of results, interdependency of tasks, decision related to the choice of software development platform, and the non-availability of data resources and/or metadata (each one of these risks is detailed in the relative WP). The Consortium, under the direction of USG, will address all these issues, whenever they arise.

Task 9.1: Scheduling and forecasting (USG). The Project Manager will plan an initial schedule and maintain this schedule using a Project Co-ordination Tool, e.g. Microsoft Project. The related files will be accessible by the Project Co-ordination Committee (PCC) members.

Task 9.2: Workpackage co-ordination (all) The Work Package Leaders will be responsible for the structure and timing of the deliverables, complying in line with management guidelines. The WL will allocate the responsibilities required to complete each deliverable between the participant partners based on their profile, experience and assigned manpower, and will identify task leaders (TLs). The WL will also co-ordinate and WP level meetings and will compile the required information from the WP participants on order to produce the final deliverables and reports for the PCC.

Task 9.3: Documentation and electronic repository (USG). All project documentation will be prepared and stored in an electronic repository in a common format that will be agreed at the kick-off meeting. All generic documents (deliverables, meeting minutes and internal reports) will be normalised to maintain homogeneity in the project. A project electronic repository will be available to the consortium members, reviewers and project officer where relevant information and shareable code of the project will be stored and updated. An additional public access site will be developed to promote the findings of the project. Confidential information will be managed through electronic mail addressed personally to the appropriate responsible people. In order to ease the management of documentation in the repository, a naming convention will be agreed at the kick-off meeting.

Task 9.4: Reporting (USG). By-monthly management reports will be prepared by the Program Manager, who will also prepare six-monthly (and annual) scientific progress reports and cost statements with inputs from the WP and Task leaders.

Task 9.5: Quality Assurance and software engineering standards (USG). A quality control policy will be defined and implemented for the project achievements. In particular, a Quality Assurance Plan to assure the quality of the project deliverables will be prepared by the PM to be approved by the PCC. Quality assurance procedures will be developed in line with relevant standards (e.g. BS5750 and ISO 9001). These will include: Appropriate software metrics (e.g. errors per KLOC, MTBF); Developer tests (e.g. module, module interaction, connectivity, performance); User acceptance tests (e.g. scenarios, functions, naïve users); Change management (e.g. version control, change analysis and prioritisation); Configuration management protocols (e.g. builds and release versions); Documentation (e.g. requirements, user manual, design documents). A collection of software engineering standards and application program interfaces will be defined and agreed in order to ensure the correct integration of any modules developed by the different partners.

Task 9.6: Monitoring processes, self-assessment and conflict resolution (USG). Special attention will be

dedicated to self-assessment at project-level and for evaluation activities. This will be done through the validation phases (verification and demonstrations), the reporting activities and the meetings of the PCC. These meetings will be maintained at each project milestone, and it will be the responsibility of the PCC to evaluate the progress towards the project's objectives, as well as to decide and manage any appropriate corrective actions. In addition, there will be an emphasis on collaborative assessment, whereby, in order to maintain standards of software performance and to improve the application's integration, partners will be promoted to test system components which they have not developed.. An Escalation Procedure will be established to resolve any conflict that could arise. This procedure will contain guidelines (to be agreed in the kick-off meeting) based on the conflict type (technical, exploitation, administrative or management). Conflicts are expected to be escalated through WLs, the PCC and finally the SC. Each representative on the SC will have a single vote.

Deliverables

D9.1: Development of web-based electronic repository to support project co-ordination. Project reports and management documentation will also be produced in line with the normal requirements of the Commission.

D9.2: Overall project management.

D9.3: Bi-monthly management reports.

D9.4: Six-monthly progress report.

D9.5: Six-monthly cost statements

D9.6: Annual reports

D9.7: Final report

Milestones and expected result

PM2: D9.1 completed.

PM36: D9.2 Overall project management completed.

PM36: D9.7 completed.

PM2, PM4, PM6, ...: D9.3 completed.

PM6, PM12, PM18, PM24, PM36: D9.4 and D9.5 completed.

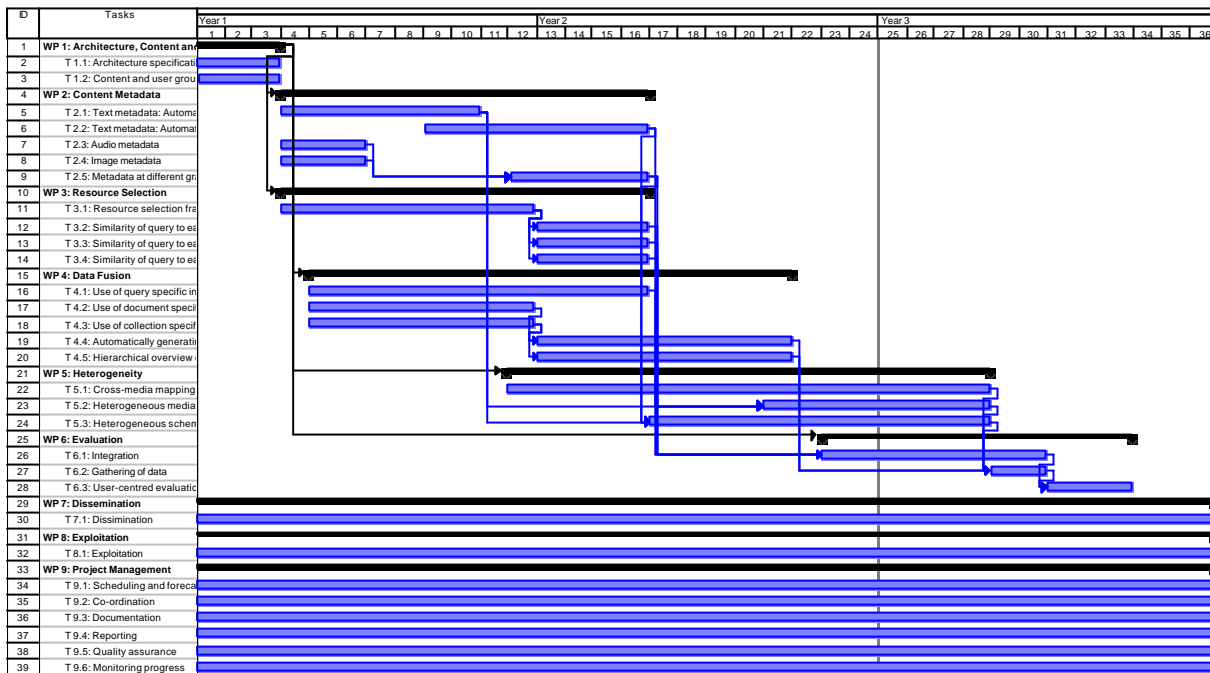
PM12, PM24, PM36: D9.6 completed

9.4 Deliverables list

Del. no.	Deliverable name	WP no.	Lead participant	Estimated person-months	Del. type*	Security**	Delivery (proj. month)
D1.1	System architecture	1	UNIDO	5	Report	INT	PM3
D1.2	Content and user groups	1	UNIDO	4 (+ 0.5)	Report	INT	PM3
D1.3	Development and evaluation plan	1	UNIDO	1	Report	INT	PM3
D2.1	Resource description formats	2	DSI	10	Report	PU	PM6
D2.2	Tools for extraction and verification of resource descriptions	2	DSI	12	Prototype	PU	PM16
D3.1	resource selection framework and methods	3	UNIDO	10	Report	PU	PM12
D3.2	Tools for resource selection	3	UNIDO	10 (+ 1)	Prototype	PU	PM16
D4.1	Methods for data fusion	4	USFD	12	Report	PU	PM12
D4.2	Tools for data fusion	4	USFD	10 (+ 1)	Prototype	PU	PM16
D4.3	Methods and tools for automatic generation of surrogates and presentation of fused results	4	USFD	10	Report Prototype	PU	PM21
D5.1	Methods for resource selection and data fusion for heterogeneous collections.	5	USFD	24 (+ 2.5)	Report	PU	PM 28
D5.2	Tools for resource description extraction, resource selection and data fusion for heterogeneous collections	5	USFD	14.7 (+ 1)	Prototype	PU	PM 28
D6.1	Component evaluation and integration	6	USG	15 (+ 2.5)	Report Prototype	PU	PM33
D6.2	Test-bed for resource selection and data fusion for multimedia Digital Libraries	6	USG	18.3 (+ 2)	Test-bed	PU	PM33
D6.3	User-centred evaluation of the MIND system	6	USG	6	Report	PU	PM33
D7.1	Web-site to support project promotion.	7	USG	1 (+ 2)	Web site	PU	PM3
D7.2	Project presentation	7	USG	(0.25)	Report	PU	PM3
D7.3	Dissemination and use plan	7	USG	(1.5)	Report	INT	PM6
D7.4	Workshop to publicise and discuss results of project.	7	USG	1	Workshop	PU	PM36
D8.1	Technology implementation plan	8	USG	1 (+ 1)	Report	INT	PM36
D8.2	Promotional brochure	8	USG	1 (+ 1.75)	Brochure	PU	PM30

D8.3	Exploitation and business plan	8	USG	(1)	Report	INT	PM36
D9.1	Web-site to support project co-ordination.	9	USG	2	Web site	INT	PM2
D9.2	Project management	9	USG	17 (+ 5.5)	Other	INT	PM36
D9.3	Bi-monthly management reports	9	USG	(1)	Report	INT	PM2, PM4, PM6, ... PM28, PM30, PM36
D9.4	Six-monthly progress report	9	USG	(1.5)	Report	INT	PM6, PM12, PM18, PM24, PM30, PM36
D9.5	Six-monthly cost statements	9	USG	(1.5)	Report	INT	PM6, PM12, PM18, PM24, PM30, PM36
D9.6	Annual reports	9	USG	(1)	Report	INT	PM12, PM24
D9.7	Final report	9	USG	1	Report	PU	PM36

9.5 Project planning and timetable



9.6 Graphical presentation of project components

The inter-relationships between the various tasks have been already reported in the GANT in section 9.5.

9.7 Project management

Structure and Co-ordination, Allocation of responsibilities

The management structure will be based on a minimal number of committees and members, with the objective of improving the overall flexibility and swiftness of the decision processes. The people responsible for the daily management of the project will be the Project Manager and the Work Package Leaders

Project Manager

The Project Manager will be responsible for the following tasks: Project scheduling and forecasting. Surveillance of the resources and work-content deviations; Overall technical management and co-ordination of the project, revision of internal reports; Production and consolidation of periodic progress reports, and co-ordination of the project final ones; Cost statements consolidation, and financial co-ordination; Management of project-level meetings; Interfacing with EC officers and external reviewers. Co-ordination of the periodic progress reviews; internal storage and dissemination of the information (communication strategy) including management of the Consortium web-site. USG as co-ordinator will assume responsibility for the project management based on its experience in previous ESPRIT and other research projects.

Work-package Leader

Each Work-package will have a Leader, nominated by the corresponding partner. This person will be responsible for its activities and deliverables (functionality and quality).

Committees

The Steering Committee (SC) will consist of one senior representative from each partner. It will be responsible of global supervision, and decisions in the event of problems at lower levels. This committee will settle any conflicting situations, which may occur during the project and which cannot be solved autonomously. USG's representative will chair the Committee.

The Project Co-ordination Committee (PCC) will consist of the Project Manager, who will chair the committee, and at least one representative from each partner (WLs and TLs). It will be responsible for periodically reviewing the technical and administrative work of the project. The PCC will hold meetings at least once every six months. It will also review exploitation opportunities for the project results. The responsible persons for each company and WP (technical, exploitation, and administrative) will be finalised at the project kick-off meeting. This information will be distributed and maintained by the Project Manager via the Consortium web-site. A Consortium Agreement clearly stating the management, responsibility and exploitation issues, including intellectual property and commercialisation rights will be signed by the partners within one month of the project starting date.

Project Co-ordination

In addition to the tasks reported in WP9, the project coordination will implement the following general procedures:

- ◆ **Meetings** - The meetings will be organised by the Project Manager or WP Leader along with the representative from the hosting organisation. This will include preparing a pre-agenda and a post-report including the meeting minutes in a pre-normalised form, for comments and approval of the attendants. If no comments are received before two weeks the meeting minutes will be considered approved.
- ◆ **Deliverables** - A set of normalised deliverables will be defined. The deliverables will be issued after being reviewed by the WP Leaders, and released by the PCC after a new QA version, the Peer Review. The Peer Reviewers will be nominated for each deliverable by the PCC and will be initially chosen from people not directly involved in the specific WP. A draft model for the PRR will be defined. All deliverables will be of a sufficient quality to ensure that, if the consortium approves, they may be used for external dissemination of the project results.

Justification of equipment needs

In the following is a summary justification of the equipment needs of the consortium partners, as envisaged at this stage and relative to the generic test-bed architecture depicted in figure 1.

USG needs a server with appropriate storage devices for making available the web-sites relative to WP7 and WP9. This server will also be used for the evaluation of the final demonstrator to be carried out in WP6 and for prototype development (in particular with audio) relative to WP2. A laptop is needed for work to be carried out in WP6 relative to the evaluation of information access using mobile computing, moreover the laptop is needed for presentation of the project results at conferences, visits to academics and business contacts. Both the laptop (used as a client) and the server will be used for all other work in which USG is involved. USG also need a PC for the project coordinator.

UNIDO needs 2 complete PCs for work to be carried out in WP1 and WP3. One of these PCs will be used as server for a Digital Library, the other will be used for prototype development.

DSI needs a specific machine configuration (SGI-O2 and Octane Duo) for work to be carried out in WP2 and WP4 on resource selection and data fusion of images/graphics. In addition these machines will be used for work of prototype development of graphical user interfaces relative to WP4 and WP5. A laptop is also needed for project management, dissemination and as a mobile resource for graphical user interface testing. Acquisition and storage devices (in particular a DVD-R) are needed for keeping local copies of Digital Libraries and for prototype development and testing (with images, in particular).

USFD needs a server for work to be carried out in WP2 and WP5. In both WPs USFD is involved in prototype development and the server will be particularly useful for work to be carried out on audio data relative to WP2 and WP3.

CMU needs a high end PC for the work on automatic acquisition and verification of resource descriptions. In addition, 1 smaller PC with a large disk will be user as server to provide a small test-bed Digital Library, and another small PC will be used to develop and test user interfaces relative to WP4 (in particular the hierarchical overview of the fused results).

All the above equipment will be fully integrated into the respective department networks enabling full connectivity of these machine for the use of the consortium as a whole.

10 Clustering

MIND will establish strong collaborative links with other EU funded DL projects. In particular with:

- DELOS NoE: Test-bed for DLs;
- CYCLADES: DLs in an cooperative environment;
- SCHEMAS - Forum for Metadata Schema Implementers;
- SCHOLNET.

Some of MIND partners are directly involved in these project, making it easier to collaborate.

In addition, MIND will establish collaborations with other nationally funded DL projects, such as:

- BUBL;
- GDL;
- SCONE;
- CAIRNS;
- Clyde Virtual University;
- GAELS;
- DAFFODIL.

11 Other contractual conditions

None.